Using Fathom® statistical education software in high school to examine students' acceptance of virtual simulation and use of simulation to model sample size when sampling from large and infinite populations

Anthony Frederick Bill
Bachelor of Engineering (Chemical), Bachelor of Teaching

Submitted in fulfilment of the requirement for the degree of PhD,
University of Tasmania
November, 2012

Declarations, Statement of Co-authorship, Abstract, Acknowledgements,
Table of Contents, List of Tables, List of Figures,

Chapters 1 – 5,

References, Acronyms & Initialisations, Glossary, and
Mathematical Symbols

Declarations

Declaration of Originality

This thesis contains no material which has been accepted for a degree or diploma by the University or other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Authority of Access

This thesis may be made available for consultation, loan, and limited copying. This permission covers only single copies made for study purposes, subject to the normal conditions of acknowledgment in accordance with the *Copyright Act*, 1968.

Statement of Ethical Conduct

The research associated with thesis was conducted in accordance with the Human Research Ethics Committee (Tasmania) Network (HERCS) (Approval No. H009790), the Department of Education Tasmania (Ref. 672670), and the Teachers Registration Board of Tasmania Code of Professional Ethics for the Teaching Profession in Tasmania. To the researcher's and the supervisors' knowledge no concerns regarding the ethical nature of the study or the personal conduct of the researcher were raised by the schools, the principals, the colleague teachers, parents or the students, HERCS or the Department of Education Tasmania.

Funding

This project was jointly funded by the Australian Research Council (ARC) through an Australian Postgraduate Award Industry (APAI) Linkage project grant (LP0669106) and by the publisher of Fathom software Key Curriculum Press of Emeryville, CA.

Signed, 16th November, 2012

Tony Bill

Anthony Frederick Bill

Statement of co-authorship

The following people contributed to the publication of the work undertaken as part of this thesis:

Bill, A. F., Henderson, S., & Penman, J. (2010). Two test items to explore high school students' beliefs of sample size when sampling from large populations. In L. Sparrow, B. Kissane & C. Hurst (Eds.), *Shaping the future of mathematics education.* (Proceedings of the thirty-third annual conference of the Mathematics Education Research Group of Australasia, Vol. 1, pp. 77–84). Fremantle, WA: MERGA. Retrieved January 12, 2011, from http://www.merga.net.au/publications/counter.php?pub=pub_conf&id=833Paper 1

Candidate/ author 1 (80%), author 2 (10%), and author 3 (10%)

Author 2 and author 3 co-contributed to the conduct of the study and reviewed the paper.

We the undersigned agree with the above stated "proportion of work undertaken" for each of the above published (or submitted) peer-reviewed manuscripts contributing to this thesis:

Signed and dated: Anthony Frederick Bill Tony Bill 27/2/12

Sally Henderson Sally Henderson 27/2/12

John Penman John Penman 27/2/2012

Abstract

Statistical literacy is regarded as essential for good citizenship, employment, and practical day-to-day living. The ubiquitous nature of data and computers in contemporary society has increased both the need for statistical literacy and the means of developing statistical literacy.

This study investigated students' acceptance of Fathom® virtual simulation and re-sampling as a legitimate mathematics tool, the teaching and learning of the explicit determination of sample size when sampling from large populations, and students' development of use of Fathom statistics education software.

The study was conducted as a three-week long classroom unit of work taught in two Year 9 classes and a detailed study of twelve students in Tasmania, Australia. Pedagogical best practice principles derived from statistics education research guided the study. These included engagement with the big ideas of statistics, active learning and data sets students can understand and value, statistical enquiry that cultivates statistical habits of mind, the use of technology tools that allows students to explore data and concepts, mathematical experiences of substance, provision of a developmental pathway for students to study statistics at more senior years, and authentic assessment.

Fathom was developed for senior high school and tertiary study, and its use in Australian high schools is relatively novel. Students' unfamiliarity with the software presented at least two challenges: developing acceptance of Fathom's virtual re-sampling probability simulator as a legitimate mathematical tool and acquiring basic fluency in the software's use such that the software was not a constraint on learning. Students' acceptance of the probability simulator was cultivated purposefully through a process of formal statistical enquiry where students examined the fairness of the Fathom virtual die. Students' development of use of Fathom re-sampling was examined from the three aspects of key terminology, graphical data representations, and their relationship with Fathom. The principles of instrumental genesis guided the introduction to, and the examination of, students' use of Fathom.

Sample size is presently ignored in the high school curriculum, and students may complete formal school education with unsophisticated notions of sample size, possibly first acquired in upper primary school. The sample size model $e = \pm 1/\sqrt{n}$, which relates the sample size n, to the margin of error e, of the accuracy of measurement, was used in this study. A foremost consideration was that the model was potentially accessible and that students could apply their understanding in a real-life context. Large populations were studied because formal mathematical treatment is relatively simple. Students' work samples were assessed using the SOLO taxonomy, and situated abstraction was used to observe students' development of understanding of selected mathematical concepts.

The study concluded that a process of statistical enquiry may be used both to promote acceptance of virtual simulation and to foster the development of statistical "habits of mind." The sample size model $e = \pm 1/\sqrt{n}$ has application in Year 9 principally to mathematise traditional Law of Large Numbers activities, where the computing power of virtual simulation allows exploration of very large sample sizes. The introduction of re-sampling and the sample size model in Year 9 provides the foundation for the consideration of contextual tasks in more senior school years. The study suggests that Fathom is suitable for Year 9 students, but recommends further research in the use of re-sampling to exploit fully the software's potential.

Acknowledgements

Many people have been influential during the research and writing of this thesis. I would like to thank the following people for their encouragement and support during this interesting, and occasionally, challenging time.

My partner Louise Oxley, without whose love, loyalty, and steadfast support this thesis would not have been completed. Thank you, my dear friend.

My daughter Emily, who, in the time of this thesis, has grown into a highly talented, intelligent, beautiful, gracious woman of wisdom and maturity.

My co-supervisor Emerita Professor Jane Watson, whose professionalism, dedication, and untiring support throughout the writing of this thesis were astonishing. Thank you Jane.

My co-supervisor Associate Professor Rosemary Callingham, whose sensible strategic advice has been of immense value.

My adviser, Mr. William Finzer, of KCP Technologies, Emeryville CA, who helped shift my gaze to the international community.

My two colleague teachers whose generosity, professionalism, and kindness are gratefully acknowledged.

I would also like to acknowledge and thank:

The students who participated in the study

Colleagues and staff at the University of Tasmania

My colleagues in the wider statistics education research community

The three industry partners The Australian Bureau of Statistics; Key Curriculum Press, USA; and the Noel Baker Centre for School Mathematics, Prince Alfred College, South Australia.

Table of Contents

Chapter 3 Methodology

Chapter 4 Results

Chapter 5 Discussion and Implications

Appendices

Appendix A Classroom study materials

Student worksheets

Lesson MS-PowerPoint presentations – Boys' class

Lesson plans – Boys' class

Tables

Tables (cont.)

Tables (cont.)

Figures

Figures (cont.)

CHAPTER 1 Introduction

1.1 Overview

This study investigated use of the statistics education software Fathom® in high school to examine students' acceptance of virtual simulation as a legitimate mathematics tool, the explicit determination of sample size when sampling from large and infinite populations, and students' development of use of Fathom. The study was conducted as a three-week long unit of work taught in two Year 9 classes – one all male and one all female – at two government funded schools in Tasmania, Australia.

Statistics education research emphasises the importance of students' beliefs related to probability and statistics, and that misconceptions students bring to the classroom may confound learning (Batanero & Diaz, 2007). The study sought to promote students' acceptance of the Fathom coin and die simulation through a purposeful approach of statistical enquiry.

The explicit determination of sample size is presently ignored in the high school curriculum, and students may complete formal school education with unsophisticated notions of sample size possibly first acquired in upper primary school. These notions may, for example, consider an appropriate sample size as a proportion, such as "10% of the population." The explicit determination of sample size is a complement to the sampling concepts of a random and representative sample that are taught presently in high schools, and it is also a response to a question posed naturally by students: "What sample size should I use?"

Fathom®, the statistics education software used to support learning, is a product of Key Curriculum Press (2005), and one of a suite of products that includes Geometer's Sketchpad® and Tinkerplots®. Fathom was developed on principles derived from education research that include the value of multiple representations of data, dynamic linkages amongst data representations, models and representations familiar to the user, simulations that draw attention to mathematical concepts, and features that allow exploration and construction of mathematical meaning. The software is intended for use by senior high school, college, and tertiary level students, and its use with the younger high school cohort in the study is relatively uncommon. The study adopted the approach that the introduction and use of the basic features of Fathom at high school provides a foundation for more intensive use of Fathom at senior school levels and above.

The study gave students their first exposure to the Fathom software. Students' unfamiliarity with the software presented at least two major challenges peculiar to the use of the software tool that were separate to students' development of understanding of the statistical concepts in the study. The first challenge students faced was to acquire efficiently basic skills and fluency in the use of the software such that the use of the software did not confound learning. The second challenge was to develop confidence in Fathom's virtual probability simulator as a legitimate mathematical tool. The study sought to address the first challenge using principles developed from instrumental genesis (e.g., Drijvers, Kiernan, & Mariotti, 2010), and to address the second challenge through a process of statistical enquiry where students purposefully explored the software's legitimacy through a comparative study of the fairness of familiar physical and virtual simulations.

The study included the use of a formal algebraic model to calculate sample size $e = \pm 1/\sqrt{n}$ (Shaughnessy & Chance, 2005) that related the sample size n, and the accuracy of the survey through the margin of error, e. In proposing a mathematical model the foremost consideration should be that the model is accessible: students can potentially both understand the key underlying concepts and apply the model in a real-life context. The large populations of national and state voting populations were studied because formal mathematical treatment is relatively simple.

Pedagogical principles considered to be best practice from statistics education research were used (e.g., Ben Zvi & Garfield, 2004). These principles included engaging students with the big ideas of statistics, statistical enquiry, classroom discussion, technology that allowed students to explore concepts, and authentic assessment (Archbald & Newman, 1988).

1.2 Statistical literacy in the high school curriculum

Statistical literacy is regarded as an essential disposition for good citizenship, for employment and professional life, and for practical day-to-day living. Statistical literacy is arguably a sub-set of quantitative literacy and the term statistical literacy is used throughout this study.

The state Department of Education, Tasmania curriculum framework emphasises a cross-curriculum approach to teaching and learning. Statistical literacy has application across the curriculum in subjects traditionally identified as Science, Studies of Society and Environment (SOSE), or Health and Physical Education (HPE).

The ubiquitous nature of data and the widespread availability of computers have created both the need for statistical literacy and the means of developing statistical literacy amongst students. The demands of society drive change in both the school curriculum and pedagogy. Educational statistics software was available and progressively more sophisticated, but it was relatively untested in Tasmanian schools. Fathom was not used presently in Tasmanian schools, but it is used extensively internationally, for example in the USA and Germany (e.g., Biehler & Prommel, 2010). These developments, in combination with the ready availability of computers and electronic data in schools, provide the justification for educational research in computer-based statistics education.

1.3 The research study as a Year 9 teaching unit

The research study was conducted with two Year 9 mathematics classes at two metropolitan Tasmanian government funded single-gender schools. Although the two classes were defined as advanced mathematics classes both colleague teachers considered their classes as capable, but of mixed ability and the students had self-selected to enrol in the course. A class of 35 female, and a class of 21 male, students participated.

The schools were selected as two government schools in Tasmania, the colleague teachers were nominated by the school as the senior mathematics teachers in the schools, and the colleague teachers nominated their Year 9 extended mathematics classes as the classes most likely to benefit from the study. The school, the colleague teachers, and the students proved very supportive of the project.

The philosophical principles underpinning the study were the importance to learning of the development of a culture of enquiry and statistical process; the belief that technology was designed to support, but was not the central element of, the learning process; a topic that was anchored to formal mathematics appropriate for the students; and a developmental pathway for students for the study of mathematics at more senior years.

Statistics education research does not occur in a vacuum: individual people are involved. From a student's perspective a research study conducted as a classroom unit of work – as this study was – must provide a learning experience at least as valuable as any other topic in the course. From colleague teacher's perspective the research had the potential to provide exposure to novel teaching opportunities and professional development, and to inform pedagogy. The participating school had the opportunity to observe and actively contribute to the development of new teaching material.

1.4 Overview of this study

This thesis comprises five chapters and appendices. A consistent format and structure is used throughout the thesis: all chapters have an introduction that provides an overview of the chapter, sub-sections that present or examine key themes, and a concluding summary.

This chapter, Chapter 1, provides an overview of the thesis.

Chapter 2, the Literature Review, examines the education research literature relating to the teaching and learning of statistics and probability in high schools. Five themes are examined: (a) theoretical frameworks used in education research, (b) statistical thinking in society and education as a justification for statistics education, (c) statistics and probability education in the classroom that includes consideration of procedural and contemporary statistics education, the cultivation of statistical thinking and class discussion, students' notions of data and data aggregates, coin and die systems, students' beliefs of chance events and sample size, measurement and error, mathematical modelling and the sample size model used in this study, and the interpretation of graphical data representations; (d) computer technology and statistics education software, theoretical models to introduce software into the classroom, and computer simulation; and (e) assessment of statistical thinking. The chapter concludes with a summary of the implications of the literature review for the study and a statement of the three research questions.

Chapter 3, the Methodology, describes the research method, the research setting and cohort, and the student work samples used. Drawing on the review of the education research literature the chapter provides the justification for the design of the research study including the methodology, the pedagogy, the topics, software, student cohort, data collection methods and assessment framework used. The study was conducted in two parts: a classroom study in two Year 9 classes taught by the researcher as a unit of work supported by colleague teachers, and a detailed study of a three-hour workshop of six student pairs who participated in the classroom study. The classroom component was conducted in four phases that broadly addressed each of the three research questions. The first phase of the classroom study provided pre-study testing of students' basic mathematical skills and beliefs of sample size, the second phase examined students' acceptance of the Fathom simulator as a legitimate mathematics tool, the third phase considered students' use of a formal large population sample size model, and the fourth and final phase provided post-study testing of students.

Chapter 4, the Results, presents the data from the assessment items from both the classroom and the detailed study with illustrative student exemplars, supported by extracts taken from interviews with the two colleague teachers and the students' post-study questionnaire. The data are presented in four sections where the first section presents the students' pre-study data followed by three sections that address each of the three research questions. Students' development of understanding was assessed using the SOLO model, students' development of use of the Fathom software and development of understanding of the mathematical concepts were examined using the instrumental genesis framework supported by aspects of situated abstraction.

Chapter 5, the Discussion and Implications, addresses the three research questions using the data collected in the study. The chapter begins with an overview of the study and considers whether the study was conducted consistent with the methodology. The three research questions are then addressed in sequence. The affordances and constraints of the pedagogy and software are identified. Implications for teachers, education researchers, software developers and teaching resource developers are discussed, and the chapter concludes with a brief summary.

CHAPTER 2 Literature Review

2.1 Introduction

This chapter examines the literature relevant to the teaching and learning of statistics and probability in the contemporary school classroom. Particular emphasis is placed on the use of computer technology and best-practice pedagogical approaches to support learning. This examination of the literature is presented in five sections: (a) the theoretical research framework used in this study; (b) probabilistic and statistical thinking and its role in decision-making in contemporary society as a justification for statistics education in schools; (c) statistics and probability education in the classroom, including pedagogical best-practice; students' notions of data, chance, and sample size; measurement, measurement error, sampling as measurement; mathematical modelling of sample size; and graphical data representations; (d) computer technology in schools, theoretical frameworks used to introduce technology, and computer simulation and re-sampling; and (e) methods of assessment of probabilistic and statistical thinking. The chapter concludes with a brief summary and a statement of the three research questions.

2.2 The theoretical framework of the study

2.2.1 Introduction

This section briefly considers world views to position the researcher's own perspective. Quantitative and qualitative research methods are examined, and the perceived limitations of the two lead to a consideration of a mixed-method approach. The Scientific Research Approach (Shavelson & Towne, 2002) is explored as a framework for education research. Research in technology in mathematics education is the subject of considerable recent criticism (e.g., Reeves, 2006) and this criticism provides an opportunity to refine the research design (e.g., Teddlie & Tashakkori, 2010), and to incorporate strategies to promote research validity for this study.

2.2.2 World-views and the researcher's world-view

One's research perspective is shaped by one's world-view: one's own personal philosophy, beliefs and assumptions (Guba & Lincoln, 1989). Creswell (2009) describes four world-views: post-positivism, social-constructivism, advocacy, and pragmatism. The post-positivists' world-view (Phillips & Barbules, 2000) is essentially scientific and deterministic where the researcher seeks to identify and assess the causes of phenomena objectively. It is a rational and dispassionate research approach. Knowledge is conjectural, imperfect and fallible, relying on data and evidence to shape knowledge and to develop relevant conclusions. A social constructivist world-view assumes that individuals seek an understanding of the world in which they live, and that these views are as varied and complex as the individuals themselves. Crotty (1998) identified three key characteristics of a social constructivist world-view: that meaning is constructed as people engage with the world, that people engage with the world based on their own historical and cultural perspectives, and that people develop meaning through a social context that arises through interaction. An advocacy world-view is highly intertwined with politics and social change and seeks to change the lives and the institutions of the participants. (Kemmis and Wilkinson, 1998 as cited in Cresswell, 2009), identified the key characteristics of the advocacy world-view as focussing on effecting change – not limited simply to understanding or describing the situation; helping individuals free themselves from constraints, whether externally or self imposed; emphasising practical, rather than theoretical outcomes; and cooperating and collaborating with participants. A pragmatic world-view focuses on "what works" utilising all available approaches and mechanisms to understand the situation. Drawing on the work of Morgan (2007) and Cherryholmes (1992), Creswell (2009) identified three key characteristics of the pragmatic world-view as not committed to any one philosophy, complete freedom to choose methods, techniques and procedures, and a focus on the research question rather than the method. Education and prior professional experiences lead the researcher to identify most naturally with the post-positivist and the pragmatic world-views.

2.2.3 Qualitative, quantitative, and mixed-method research approaches

Qualitative research is a "situated activity [that] consists of a set of interpretative material practices that make the world visible" (Denzin & Lincoln, 2003, p. 4). Creswell (2003, 2009) and Miles and Huberman (1994) identified collectively seven defining characteristics of qualitative research: (a) naturalistic, which enables the researcher to be involved in the experiences and the world of the participants, (b) interactive and humanistic, which allows active participation and the development of rapport and credibility with the participants; (c) emergent, rather than tightly prefigured, which provides flexibility during the study; (d) interpretative and descriptive; (e) holistic, which seeks to capture complexity and interactivity; (f) introspective, which recognises the researchers own influences on the study – a feature particularly relevant where the researcher acts as participant-observer; (g) iterative and simultaneous, with a process that cycles back and forth though data analysis, design, and problem reformulation. Multiple strategies are employed, and the approach commonly uses data sources of text and image (Creswell, 2003), interviews, and observational field notes. The emphasis is on descriptive material.

Quantitative research "emphasises the measurement and analysis of causal relationships" (Denzin & Lincoln, p. 13), an approach where variables are "measured or assessed with respect to or on the basis of quantity; that may be expressed in terms of quantity; quantifiable" (Oxford English Dictionary, 1989). This is a research method of the positivist tradition where objective, value-free evidence can reveal the underlying truth. The research approach commonly uses data collection strategies of surveys and experiments with tightly controlled variables to test hypotheses; in short, but not exclusively, it has an emphasis on numerical data.

Stake (1995) considers the fundamental difference between qualitative and quantitative research as not so much one of style as of objective: qualitative provides observation that seeks an understanding of the "the complex inter-relationships among all that exist" (p. 37), and quantitative seeks to identify cause, explain, and control.

The two methods have formerly been the centre of debate, over an extended period, of the relative merits of the two approaches (e.g., Howe, 1992; Smith & Heshusius, 1986). More recently this debate has led to polarisation that is seen as entirely divisive and unproductive, and driven in part by misconceptions (Miles & Huberman, 1994; Ercikan & Roth, 2006; Onwuegbuzie & Leech, 2005). This debate does, however, serve to highlight the limitations of the two research approaches. Purely qualitative research is criticised as being unscientific, exploratory, or subjective (Denzin & Lincoln, 2003), or lacking in rigour (Onwuegbuzie & Leech, 2005), or yielding evidence that cannot be verified or generalised (Ercikan & Roth, 2006). Purely quantitative research is criticised as inadequately capturing human experience. This debate led Creswell (2009) to consider the similarities shared by quantitative and qualitative research approaches. Both strategies examine phenomena systematically and coherently, use observations to address research questions, include safeguards to minimise bias and invalidity, use analytical and data reduction techniques designed to extract the most meaning from the data, and provide descriptions that interpret the data.

The multi-method approach recognises the limitations of the pure quantitative and qualitative approaches by combining methods from both techniques (Johnson, Onwuegbuzie, & Turner, 2007). The multi-method approach combines qualitative approaches of interviews and observations, with quantitative research approaches of formal testing and surveys. Integrating the two approaches is designed to provide breadth and depth of understanding and corroboration (Johnson et al., 2007), to describe systems that have inherent quantitative and qualitative aspects (Ercikan & Roth, 2006; Teddlie & Tashakkori, 2010), and to give the best understanding of a complex situation by using the strengths of both approaches (Croninger & Valli, 2009). The two approaches are seen as complementary (Miles & Huberman, 1994). Despite the apparent advantages of the mixed method approach a review of the methods used in mathematics education research only 29% of 710 studies examined used the mixed method approach (Hart, Smith, Swars, & Smith, 2009).

Mixed-method research is imagined as lying on a continuum between quantitative and qualitative research (Creswell, 2003) with the centre of the continuum – the point where qualitative and quantitative methods are accorded equal status – as locating mixed-method research. Granting equal status suggests an attitude of mind of the researcher, rather than collecting equivalent quantities or quality of data. Mixed-method is not simply a research method, but more importantly a way of thinking (Greene, Benjamin, Goodyear, 2001;; Teddlie & Tashakkori, 2010).

Given the various short-comings of the various research methods, Creswell (2003) provided three broad criteria for selecting a research design for the research problem: the personal experience of the researcher, the style of the researcher, and the target audience. This would suggest a research objective that can be generalised to other audiences, albeit in a limited fashion, but still provide a rich view of individual experiences; a researcher who can utilise effectively systematic quantitative methods and the flexibility and exploratory nature of qualitative methods, and an audience that will accept a mixed method approach.

2.2.4 Scientific Research Approach (SRA)

Fitzallen and Brown (2007), in a review of the research model literature, categorised education research as either scientific experimental design research or correlative and descriptive research. In identifying the short-comings of existing research models Fitzallen and Brown turned to the Scientific Research Approach advocated by the National Research Council [NRC] (Shavelson & Towne, 2002). The NRC considered the nature of progress in science, and concluded that science was a statement of theories and models that are capable of rigourous hypothesis testing, scrutiny, and review based on evidence. According to the NCR:

> Knowledge is generated through a sequence of interrelated descriptive and causal studies through a constant process of refining theory and knowledge. These lines of enquiry typically require a range of methods and approaches to subject theories and conjectures to scrutiny from several perspectives (p. 123).

The principles of scientific research are applied to education research, with the term principles chosen purposefully as sufficiently flexible to be adaptable to the task and situation. Six principles are identified as guiding all scientific research, including education research (Shavelson & Towne, 2002, p. 26):

- Pose significant questions that may be investigated empirically.
- Link research to relevant theory.
- Use methods that permit direct investigation of the question.
- Provide a coherent and explicit chain of reasoning.
- Replicate and generalise across studies.
- Disclose (publish) research.

It is an approach that is consistent with calls for the use of evidence-based teaching approaches (e.g., Slavin, 2002; Rowe, 2007; Department of Education, Tasmania, 2010). Great importance is given to the role that the research community – by implication researchers, teaching professionals, students, and education policy makers – plays in fostering the development of knowledge. No distinction is made between quantitative and qualitative research methods, or between pure and applied scientific research. It is a pragmatic approach to research design with “the research question driving the design, not vice-versa” (p. 99).

A feature of the Scientific Research Approach is an extended duration and large scale design, and this may be incompatible with the time-frame of doctoral research. Reeves (2006) gave three exemplars of doctoral technology based design research as evidence that the research design is feasible for dissertations. None of the three examine mathematics education specifically, but they share the four common features of (a) working with teacher collaborators to develop a model, (b) using authentic activities to test the model in multiple contexts, (c) demonstrating change at a local level, and (d) contributing to the development of theory.

2.2.5 Criticisms of education technology research

The deficiencies of education research are identified as having little effect on educational practice (Fishman, Marx, Blumenfield, Krajcik, & Soloway, 2004; Reeves, 2006) and possessing a credibility gap (Levin & O'Donnell, 1999), being detached from practice (Lagemann & Shulman, 1999), not addressing adequately issues of "usability, scalability and sustainability" (Fishman et al., p. 43), and not being disseminated beyond the context of its initial use (Seeto & Herrington, 2006). Reeves (1993, as cited in Reeves, 2006) identified the contributing factors to poor research as inadequate literature reviews, a lack of linkages to existing education theory, poor treatment implementation, small sample sizes, and meaningless discussion of results. Two good exemplars of research in a technology-enhanced learning environment are identified, and the design features common to both include a preliminary research phase to develop a proto-theory of research (Wang & Hannafin, 2005), a collaborative approach conducted in the school environment with teaching professionals, trials in multiple contexts in the same environment, a multiple method approach, and demonstrated change at the local school level (Reeves, 1995, as cited in Reeves, 2006). More recently Reeves remains critical of existing research calling for "socially responsible research that will lead to effective use of technology in schools" (Reeves, 2006, p. 64); such criticism mirrors the more general call for educational research to be more accountable and transparent (Walshaw, 2007).

In response to this criticism several researchers (Chatterji, 2004; Burke Johnson, 2009; Croninger & Valli, 2009; Teddlie & Tashakkori, 2009) sought to focus on the utilitarian benefits of research (Olson, 2004; Burke Johnson, 2009) and to identify the pragmatic, or specifically "what works" in education research design – essentially the research question is more important than the research methods used (Mertens, 2010). The researchers are critical of what they consider strongly quantitative and literal interpretations of scientific education research (Howe, 2009). The researchers noted that the Scientific Research Approach specifically identifies principles of research, and that the approach recommends the use of methods "that permit direct investigation of the question" (Shavelson & Towne, 2002, p. 62), rather than prescribing any specific data collection technique. They advocate a broad interpretation of scientific design, multi-methods of data collection, and extended term studies.

2.2.6 Research validity and triangulation

Quantitative and qualitative research approaches seek to address issues of data quality, but the two approaches use different terms to describe similar, often seemingly indistinguishable, concepts. In quantitative studies the terms used are validity, reliability, generalisability (Creswell 2003; Kvale, 1996), and objectivity (Creswell, 2009)*;* and in qualitative studies the parallel terms used are credibility and confirmability (Mertens, 2010, Guba & Lincoln, 1989). Validity refers to the accuracy and credibility of the research findings, and whether the study investigates what is intended to be investigated. In qualitative research validity necessitates strategies that include using triangulation, participant review, rich description to convey the findings, inclusion of counter or contrary information, prolonged time in the field, peer debriefing, and external audits. In quantitative research validity or confirmability refers to whether the instruments used (survey tools) allow inferences to be drawn. Validity can be prejudiced by inadequate instruments (e.g., a questionnaire item is ambiguous), varying the procedures or treatments (an inherent difficulty when two or more classes of students are used), and inappropriate use of statistical methods (Kvale, 1996). Reliability or dependability considers the consistency of responses to the research protocol. Generalisability refers to the transfer or application of the results to new settings or people, which can be addressed by providing rich description of the context and setting that allows readers to make their own assessment of whether the information is applicable to their situations (Mertens). Objectivity, also known as confirmability, means the conclusions drawn from the study are evidence based and explicitly linked to the data by a "chain of evidence" (Shavelson & Towne, 2002, p. 4). Objectivity, the hallmark of the post-positivist rational approach, may not be purely objective. This has led some researchers to reflect on the nature of reality (e.g., Denzin & Lincoln, 2003), with Creswell (2003) stating that one objective reality does not exist, and Denzin and Lincoln noting that there are no objective observations.

Triangulation is a process where a variety of research methods is used to enhance the accuracy and credibility, corroborate (Creswell, 2009), validate (Richardson, 2003), or minimise misperception (Stake, 1995) of research findings. Methodological triangulation, where multiple methods are used, becomes indistinguishable from a mixed-method approach (Creswell, 2009). Richardson criticises triangulation as an inadequate metaphor, and one that implies the existence of a single fixed point of reality. Validation is re-imagined as the more complex process of crystallisation*,* which is a process designed to clarify progressively to provide a rich understanding of the topic under study.

Formal protocols may enhance research validity, but research validity is also one of personal position and belief. Creswell recognised the inherent limitations of research by arguing that all social research is interpretative, and that the researcher should be self-reflective and mindful of how the researcher’s own background might affect the interpretation. Similarly, Freebody (2003) speaks of “the willing adoption of responsibilities” and to avoid criticism the onus lies on researchers to be “more objective, more empirical, and more rigorous” (p. 69).

2.2.7 Section summary and implications for this study

The Scientific Research Approach provides research design principles derived from scientific research that emphasise evidence-based empirical enquiry. The method incorporates the development and refinement of theories, but the term theory seems comprehensive, so in this study the alternative, and more modest terms of proto-theories (Wang & Hannafin, 2005) or teaching principles (Croninger, Buese, & Larson, 2006) are used. Critics of education research note the limited impact that research has had on professional teaching practice and argue for a pragmatic approach that will lead to sustainable change in education. The review of the literature of education research helps position this research study.

- Use a multiple-method approach to give rich description of complex situations (Croningen & Valli, 2009) giving equal weight to complementary quantitative and qualitative methods (Miles & Huberman, 1994);
- Apply the principles a Scientific Research Design (Shavelson & Towne, 2002), because the approach provides a clearly articulated research structure;
- Adopt a pragmatic (Teddlie & Tashakkori, 2010) and utilitarian approach. Generate research based useable knowledge (Olson, 2004; Burke Johnson, 2009) for the key-stakeholders of the students participating in the study, practising teachers, and policy developers;
- Contribute to theoretical knowledge within the modest objective of development of teaching principles or proto-theories appropriate to both the scope and the context of the research study (Croninger, Buese, & Larson, 2006);
- Be mindful of the cultural aspects of learning (Denzin & Lincoln, 2003), the limitations and subjective nature of research (e.g., Cresswell, 2003) , and the criticism of contemporary education research (e.g., Reeves, 2006); and
- Incorporate into the research design a collaborative approach with practising professionals (Reeves, 2006), and strategies to enhance data validity (e.g., Cresswell, 2009).

2.3 Statistical thinking in society and education

2.3.1 Introduction

This section examines briefly the importance of statistical literacy in contemporary society as a justification for its incorporation in the school curriculum. What it means to be statistically literate is considered from the perspectives of education research, professional statisticians, and the school curriculum. A distinction is made between statistical and mathematical thinking. Statistical education research identifies the big ideas of statistics as key elements of contemporary statistics education theory – ideas that may have been at some variance to the Tasmanian high school curriculum at the time of the study.

2.3.2 Statistical literacy: its importance in society and this study's definition

Statistical literacy, in its broadest sense (OECD, 2003; Steen, 1999; Wilkins, 2000), is essential for modern life: for employment and professional work, for an individual's role as an informed citizen, and for the conduct of the practical day-to-day affairs of life. Steen made an analogy with the introduction of the printing press when the written word became accessible to more than the educated elite: the relatively recent widespread availability of personal computers and the Internet may place numeric data and information at an equivalent moment in history. The sheer quantity of information, a "data-drenched society" (p. 9), contributes to the challenges of analysis and interpretation. The emphasis must be on "producing intelligent citizens who can reason with statistical ideas and make sense of statistical information" (Steen, p. 9) if they are to be "informed citizens and employees" (Gal, 2002, p. 1). In short: society demands citizens who are statistically literate.

Gal defined statistical literacy as "the knowledge and dispositions to critically evaluate and [...] express [...] opinions regarding statistical information or data-related argument" (Gal, 2002, p. 2). The models of literacy developed by Watson (1997), and separately by Gal (1998, 2002), identified knowledge and dispositional elements, which incorporated general literacy, statistical knowledge, mathematical knowledge, critical questioning, and personal dispositions of an ability to adopt a critical stance and attitudes that include a belief in the power of statistical process. Wild and Pfannkuch (1999) identified four features of objective collection of information; transnumeration or changing the representation of the data to "…facilitate understanding…" (p. 227); recognising that learning and decision-making occur under uncertainty; using statistical thinking frameworks; and synthesising contextual and statistical knowledge. Wild and Pfannkuch (1999) also described statisticians' dispositions of curiosity, awareness, scepticism, being logical and a desire to seek a deeper meaning. These are is similar to the Thinking Habits developed by Ritchhart (2001) at Project Zero and an element of the Essential Learnings Framework (Department of Education Tasmania, 2003) that was the curriculum framework at the time of the study. In this study the term statistical literacy is used in a sense that embraces both knowledge and personal dispositions.

2.3.3 Statistical thinking differs from mathematical thinking

Statistics is presently taught within the mathematics curriculum (e.g., Department of Education, Tasmania, 2010), but some researchers (e.g., Ben-Zvi, 2000; Scheaffer, 2006) argued that mathematics and statistics are two separate fields of enquiry, each with its own cultural mores. Statistics uses mathematical techniques to ensure statistical processes are rigorous and valid, in a manner similar to physics, chemistry, and economics where mathematics is used as an essential tool. In the classroom Ben-Zvi (2000) recommended that statistics should be taught as part of the social sciences, liberal arts, or business, rather than in the mathematics curriculum.

Mathematics and statistics share numbers, but in statistics numbers exist in context (Rossman, Chance, & Medina, 2006) and not in their own right. Mathematics is about proof and abstraction (Scheaffer, 2006). The value of statistics lies not in that it can be proved theoretically but in its utility. Statistics is a methodological discipline, where data are "numbers with context" (Cobb & Moore, 1997, p. 801), whereas mathematics can exist in its own right without considering potential applications. If statistical methods fail through the test of experience or application they are discarded. The emphasis on the ability to communicate results (Rossman et al., 2006) is greater in statistics than in mathematics. Mathematics is often symbolic; in statistics the verbal interpretation and explanation are central.

The need for statistics as a discipline (Cobb & Moore, 1997; Franklin & Garfield, 2006) arose because of the presence of variability and the need to quantify, understand, and explain variability. Statistics and probability are inextricably intertwined, and a study of probability provides the mathematical underpinnings of statistics. The origins of statistics lie with probability and games of chance, insurance, and interpretation of data. Statistics develops techniques and procedures to make decisions probabilistically, rather than deterministically (Scheaffer, 2006). Analysis of data may lead to differing, quite valid and reasonable conclusions, but the conclusions remain uncertain. Conclusions are often qualified, for example with expressions such as "the data suggest."

2.3.4 Education research and the big ideas of probability and statistics

If Wild and Pfannkuch (1999) and others (e.g., Gal, 2002; Watson, 1997) have identified the processes and dispositions of statistical analysis, other researchers (e.g., Garfield & Ben-Zvi, 2004) discuss the importance of the "…the big ideas of statistics…" (p. 400). Of the seven big ideas of statistics and probability – data, trend, variability, models, association, samples and sampling, and inference – five topics were incorporated into this research study: data, variability, models, samples and sampling, and inference.

Variability is the centre-piece of statistical analysis (Utts, 2003; Wild & Pfannkuch, 1999). Variation is omnipresent and natural. It has the practical consequence of making the results of actions unpredictable, but statistics helps makes sense of this complex world. A person's response might be to ignore variation – an approach that works in many situations. A second approach accommodates the variation – for example, setting an acceptable tolerance range. This study sought to consider the second approach by determining an acceptable tolerance range to a statistic by determining sample size.

2.3.5 Curriculum frameworks: sample size is missing

The curriculum framework for this study was provided by the National Council of Teachers of Mathematics [NCTM] (2000), the Australian Education Council (1994), and the Department of Education, Tasmania (2003) Essential Learnings program. The Essential Learnings Being Numerate Support Material – Teaching Emphasis (Department of Education, Tasmania, 2008) for chance and data at Standard 4 and 5, which were the levels expected of Year 9 students in the study, recommended including activities that:

- Create opportunities for calculating and using distribution summaries in the analysis and interpretation of data;
- Develop the vocabulary used to describe the analysis and interpretation of data;
- Use simulations to create a distribution with certain parameters;
- Provide opportunities to practice writing reports on inferences drawn from different types of data and justify support for hypotheses;
- Compare real and theoretical probabilities;
- Use appropriate data simulation models to devise and choose valid simulations in order to collect data when the required information cannot be collected directly;

- Allow working with "real" data sets that are a truer reflection of authentic contexts with attendant variation; and
- Actively engage in the analysis of data sets to calculate and interpret measures of centre and spread to draw conclusions.

Within Tasmanian schools the importance of representative and random sampling, and the significance of bias and sources of variation are emphasised, but the curriculum is silent on the complementary topic of explicitly quantifying sample size (Department of Education, Tasmania, 2008). Contemporary statistics education research literature also gives scant regard to quantifying sample size, with the topic not mentioned in a recent review by Shaughnessy (2007). Explicit consideration of sample size was formerly part of the senior high school curriculum (e.g., Harding, 1992), and it is an element of Level C of the Guidelines for Instruction and Assessment (GAISE) developed recently for the American Statistical Association (Franklin, Kader, Mewborn, Moreno, Peck et al., 2007, p. 69). Sample size is considered only when a survey is conducted beyond the confines of the classroom, and the practical difficulties for a teacher may discourage this type of investigation. When the opportunity for designing a survey, as distinct from a census, arises, students naturally ask the question "how many should be asked?"

2.3.6 Section summary and implications for this study

The justification for the study of statistics in school lies with the perceived skills of citizens thought necessary for effective membership and decision-making in society. Statistical literacy includes the personal skills and attributes to make decisions based on probabilistic rather than deterministic information. Statistical literacy is considered from the perspective of models developed by statistics education researchers, the distinction between mathematics and statistics, and the attitudes, skills and beliefs of professional statisticians. Of the seven "big statistical ideas" identified by statistics education researchers five are incorporated into this study. The high school curriculum at the time of the study did not include the explicit determination of sample size and the topic was not included in recent reviews of statistics and probability education (Shaughnessy, 2007), but sample size was formerly part of the senior school curriculum and it is an element of Level C of GAISE (Franklin et al., 2007).

2.4 Statistics and probability education in the classroom

2.4.1 Introduction

Earlier approaches to statistics education that emphasised a procedural approach may not promote learning effectively (Mokros & Russell, 1995). Contemporary best practice statistics education emphasises engaging students with the concepts of the "big ideas of statistics" (e.g., Ben-Zvi, 2000), developing sound intuitions and beliefs, encouraging enquiry and statistical process, and a pedagogy that includes creating opportunities for active learning, utilising electronic technology effectively, and providing authentic assessment. The importance of enculturation into statistical practice through cultivation of "habits of mind" (Chance, 2002) and classroom discourse is emphasised. Students' notions of data and data distributions, chance events, coin and die systems, and their naive notions of explicitly quantifying sample size are considered in turn. Measurement and measurement error and uncertainty associated with familiar physical measurement provide a basis for reconceptualising sampling as measurement that has inherent uncertainty. The role of mathematical modelling in schools and the limitations of existing sampling models lead to the development of three criteria for mathematical models and a sample size model that relate sample size and the uncertainty of measurement. Graphical representations are used widely to represent statistical data, so the process of, and the key factors influencing, graph comprehension are examined.

2.4.2 Procedural approaches to statistics education and their limitations

Earlier approaches to statistics education presented statistics as a collection of rules and techniques rather than a process of quantitative reasoning, problem solving or developing intuitions (Bakker, 2004; Chance, 2002; delMas & Liu, 2005; Garfield & Ben-Zvi, 2004; Pfannkuch & Wild, 2004; Scheaffer, 2006). Mokros and Russell (1995) argued that these earlier teaching approaches actively interfered with students' natural intuitive sense of basic statistical concepts, and similarly, Garfield and Ben-Zvi (2004) found traditional teaching approaches obscured the big ideas of statistics.

Garfield and Ben-Zvi (2004) and Bakker (2004) observed that although students may be able to calculate basic statistics, a sound understanding of what was being constructed or how statistical concepts were interrelated was rare; students could not generally apply the techniques sensibly. Curriculum documents (Shaughnessy, 2006) over-emphasised measures of centre, such as mean and median, and gave scant regard to variability. Assessment of statistics, according to Garfield (2003), tended to focus on the accuracy of computations, correct application of formulas, and accuracy of graphs, which provided limited information on the students' statistical reasoning and their ability to interpret statistical arguments. Jones and Tarr (2007), in review of middle-school textbooks used at the time of the study, concluded that tasks had an inappropriately low cognitive demand.

2.4.3 Statistics education – what education research considers best-practice

Current statistics education research advocates a shift from pedagogical approaches that emphasise techniques and procedures to the development of fundamental statistical thinking and reasoning skills. This latter approach, which provides the pedagogical framework for this study presented here, emphasises:

- Engaging students with data and concepts, the big ideas of statistics such as samples, sampling, variation and distribution (Ben-Zvi, 2000; Franklin & Garfield, 2006).
- Cultivating accurate statistical intuitions, conceptual understanding, and beliefs (Kadar & Perry, 2006; Pfannkuch & Wild, 2004; Watson, 2006) that support the learning of more advanced statistical concepts at senior school level (e.g., Pfannkuch, 2008).
- Active learning (Ben-Zvi, 2000; Ben-Zvi & Arcavi; 2001; Franklin & Garfield, 2006) and authentic data analysis (Groth, 2006) with real data sets and meaningful tasks in a context that students can understand and value.
- Developing a culture of enquiry and statistical process (Ben-Zvi, 2000; Franklin & Garfield, 2006), statistical thinking and reasoning (Ben-Zvi & Garfield, 2004), that uses whole class discussion where students must construct arguments and justify their positions (Groth, 2006; McClain & Cobb, 2001; Pfannkuch, 2008; Sherin, 2002; Walshaw & Anthony, 2008), and that explores both structured and unstructured activities (Chance, delMas & Garfield, 2004).
- Using technology tools that encourages students to visualise and explore data by providing different representations of the same data set (Ben-Zvi, 2000; Franklin & Garfield, 2006; Garfield, 1995; Graham & Thomas, 2005) and that

allows students to move back-and-forth between the various representations of the data (Bakker & Gravemeijer, 2004).

- Assessment that genuinely measures student learning and development (Ben-Zvi, 2000; Chance et al., 2004; Franklin & Garfield, 2006) and that accurately conveys to the student what is important (Garfield, 1995; Pfannkuch, 2005).

These principles are consistent with general principles of effective mathematics pedagogy of cultivating relationships and building a collaborative classroom community, shaping students' mathematical language, providing appropriately challenging tasks, judiciously using tools to organise mathematical thinking, and attending to the mathematics (Anthony & Walshaw, 2007).

2.4.4 Cultivating statistical thinking, discourse, and culture in the classroom

Many researchers (Ben-Zvi, 2004a; Chance, 2002; Cobb & McClain, 2004) emphasise the importance of establishing and developing what is described as socio-mathematical norms – in this instance a classroom culture of statistical enquiry – as a key element in the development of statistical thinking. This socio-cultural perspective has its origins in the work of Vygotsky (1978). Bakker and Gravemeijer (2004) noted that establishing these norms in the class is as important as computer tools, activities and the skill of the teacher, and Resnick argues that becoming competent in statistics "may be as much a matter of acquiring habits and dispositions of interpretation and sense making as acquiring skills, strategies and knowledge" (cited in Ben-Zvi, 2004a, p. 42). Ben-Zvi and Friedlander (1997) developed research programmes designed to introduce students to the art and culture of exploratory data analysis, and Ben-Zvi (2004b) wrote subsequently of the socialisation process into the culture and values of mathematics. This process of enculturation (Ben-Zvi & Arcavi, 2001; Ben-Zvi, 2004b; Biehler, 2003), has two components: development of skills and procedures and the adoption, through an apprenticeship, of a point-of-view of a community of experts.

Chance (2002) noted the importance of developing the habits of statistical thinking, but argued that the mental habits and problem solving skills needed to think statistically should be taught deliberately. It should not, argued Chance, be assumed that students would naturally develop the habits through a statistics course that does not purposefully cultivate statistically thinking. An essential feature is that the habits are established and cultivated by demonstrated example and repeated use. Such habits are an element of dispositions more generally described as intellectual character by Ritchhart (2001, 2002), and more specifically as part of statistical literacy (Gal, 2002, 2005; Watson, 1997, 2006). These dispositions can be cultivated purposefully by, for example, the *Thinking Habits* program (Ritchhart, 2002), and are now formally incorporated into some school curricula (e.g., Department of Education, Tasmania, 2008).

Differences in attitudes to learning and statistics may exist between girls and boys, and that these differences may affect the approach used. Carmichael and Hay (2009), in a study of gender difference in Australian middle-school students' attitudes to statistical literacy, found evidence that girls favoured tasks related to surveys and that boys favoured working on problems and sports contexts. The researchers also found no difference between girls' and boys' attitudes to statistical literacy, but that boys were more focussed on outcomes and that girls were more interested in mastering the process. The researchers attributed these and other observed differences, to girls' greater interest in self and their connections with real world (Powell, 2004), but they found boys' attitudes more difficult to explain. Their research has clear implications for the design of this study: offer boys an active learning environment with small learning tasks, and offer girls tasks with strong social contexts.

Classroom discourse supports this enculturation process (Walshaw & Anthony, 2008). Sherin (2002) wrote of the importance of the role of the teacher in facilitating classroom discourse and of the tension between encouraging discourse and developing significant mathematical content. The teacher acts as a "filter" (p. 205) extracting and re-focusing the discussion on the mathematical concepts. A significant barrier to discussion is that students may lack an adequate statistical vocabulary, and even in a classroom culture that encourages discussion students may not have the vocabulary to express statistical opinions confidently. Teachers may need to provide students with a working statistical vocabulary. Bakker and Gravemeijer (2004) recommend that students be allowed to use formal statistical terms loosely, and encouraged to use informal terms – for example, to describe distributions as spread out, flatter, or clumped. They claimed statistical terms would be used with greater precision as students' statistical sense develops, and that a lack of vocabulary, should not be a barrier to cultivating statistical reasoning and an intuitive sense of statistics. Language also serves to reveal thought (e.g., Godino & Batanero (1999), cited in Batanero & Diaz, 2007), and a goal of mathematics education research is developing an understanding of how mathematical meaning is constructed and evolves through instruction (Batanero & Diaz, 2007).

In considering how students developed understanding of mathematical terms Meyer (2009) adopted the approach of Wittgenstein. To paraphrase Wittgenstein's approach, words have neither a consistent meaning nor an objective meaning, and may have a multiplicity of meanings depending on the context. The simple expression of a word does not constitute meaning. It is a social constructivist approach where words acquire meaning in a social context through "language games" (Meyer, p. 905); this may be a language game of the everyday or a language game of the mathematics classroom. Two levels of understanding of a word were identified: (a) exemplaric, a lower-level of understanding where the meaning is described by example, and (b) definitional, which is a higher-level of understanding where the meaning is described by the underlying principle (Meyer, p. 910). A definitional understanding allows the meaning to be transferred to unfamiliar contexts.

Croninger, Buese and Larson (2006) noted the importance of classroom discussion in supporting learning, but refine it further after observing that high-poverty classes (and by implication less capable students) were "more dependent on the teacher to mediate the curriculum and provide multiple representations of mathematics" (p. 32). Their observation was tentatively explained, in part, by "students lacking mathematical vocabulary and basic conceptual knowledge" (p. 32). In contrast, students in high achieving classes tended to be independent learners capable of learning productively through teaching that used textbooks and worksheets. This suggests that classroom discussion may be more important for low, rather than high, achieving students. Other research (e.g., Lubienski, 2000) on the link between socio-economic status (SES) and classroom discussion reported contrary findings with lower SES students tending to be confounded and confused by whole-class discussions.

Sherin (2002) observed the challenges that teachers face in promoting class discussion and achieving a balance between the process of whole-class discussion and the mathematical content of the discussions. Other researchers (Silver & Smith, 1996) noted the difficulties of fostering student participation in discussions and that cultivating an environment of trust and mutual respect was critical. McCrone's (2005) examination of a fifth-grade mathematics class observed the evolutionary nature of the development of a culture of classroom discussion, where development was seen as a shared responsibility. Students' participation was encouraged by sharing unfinished ideas. The relationship between gender and whole-class discussion in the mathematics classroom has not been an extensive topic of research. One such study of a co-educational class of Year 6 students found that girls contributed less than boys in whole-class discussions, small group discussions are more equitable, and student-developed guidelines enhanced discussions (Theberge, 1994).

A discussion needs a topic, or a point of shared interest, on which the discussion can occur. To describe this point of shared interest Hoyles, Noss and Kent (2004) have taken from the researcher Star (1989) the notion of boundary object. In this study examples of a boundary object were a Fathom simulation and a graph of data displayed at the front of a classroom. A boundary object "provides a generalised mechanism for meanings constructed between communities" (cited in Hoyles et al., 2004, p. 320); or, alternatively, a boundary object provides shared common ground within a community of students and teacher around which mathematical meaning might be discussed, negotiated, and mediated, and where mathematical meaning may be developed through mutual construction, interaction and feedback.

Current statistical education research emphasises the importance of purposefully cultivating statistical thinking and classroom discussion to help support learning, and these consequently form an integral part of the study's pedagogical design. This pedagogy should recognise the contextual meaning of words and support students' development of a statistical vocabulary to more formal language. Such classroom discussions can occur meaningfully around the notion of a shared boundary object.

2.4.5 Students' notions of data and data distributions

Data and variability are two of the "big ideas" of statistics. Distribution is a graphical representation of the data aggregate showing the variability of the data. Konold and Higgins (2002) (cited in Hammerman & Rubin, 2004, p. 20) described four aspects of students' thinking about data. Students saw data:

- as a pointer: a focus on the data collection event rather than the actual data generated, e.g., we measured the height of everyone in the class;
- as an individual case: e.g., my height was 145 cm, the tallest in the class was 175 cm;

- as a classifier: a focus on frequencies, e.g., there were more medium height people than tall people; and
- as an aggregate: a focus on the whole data-set, such as describing a range around some measure of centre, e.g., the height of students in our class lay between 145 and 150 cm.

More sophisticated levels of statistical analysis would be demonstrated by an ability to consider the data as an aggregate. Studies of middle high school students (Ben-Zvi, 2004b; Chance et al., 2004; Hancock, Kaput, & Goldsmith, 1992; McClain & Cobb, 2001) found that students tended to perceive data as a collection of individual points rather than an aggregate; they did not demonstrate an overall sense of the data but focused on individual cases. An aggregate or global sense of the data – as shown by distribution and variability – revealed information that an individual case could not demonstrate.

Other researchers (Hammerman & Rubin, 2004; Konold, Pollatsek, Well, & Gagnon, 1996) thought that what makes data analysis complex was the need for an individual to attend simultaneously to individual values and aggregate properties. Higher level statistical thinking, according to Ben-Zvi (2004a), would be demonstrated by an ability to move between the individual and aggregate view of the data. An individual may also find a task complex if unable to adopt a global perspective of the data. Bakker and Gravemeijer (2004) recommend that formal measures of distribution such as median and quartile should be deferred until intuitive notions have been developed. Intuitive terms may provide an alternative, and less formal, way of describing distribution. Propensity was defined by Konold et al. (1996) as an intensity or rate of occurrence of some characteristic within a group. Propensity is a group tendency, as distinct from an individual's attributes, within the data set. Cultivating a sense of the propensity of a group – a fundamental but non-statistical measure – might encourage a transition from thinking primarily of individual cases to a global or aggregate view of the data. In this study the more informal term "most data" was used to describe where the range that the majority of data occurred in the distribution.

Rubin and Hammerman (2006) believe that developing an aggregate view of data may be difficult because it requires application of multiplicative reasoning as distinct from additive reasoning. Multiplicative thinking is best distinguished (Cobb, 1999) from additive thinking by example. Additive thinking would be demonstrated by partitioning a data set and then reasoning about the actual number or frequency of data within the subset; multiplicative reasoning would be demonstrated by considering the partitioned data subset as a proportion of the whole set. If the subset is a representative proportion, then inferences may be drawn of the entire population – a fundamental concept in statistics.

Multiplicative reasoning is considered by many researchers (Cobb, 1999; McClain & Cobb, 2001; Shaughnessy, 2006) as pivotal in developing statistical intuition and an ability to consider a data set as an aggregate or distribution. Cobb thought that the shift to multiplicative thinking could be supported by use of two or more data-sets with unequal numbers of data points. Additive thinking is sufficient when the two sample sets contain an equal number of data points, but a comparison based on number when the number in each data set is not equal is incorrect. This marks a crucial shift in thinking to the more sophisticated multiplicative thinking.

Distributional thinking is the ability to integrate aspects of the variability of the data aggregate, such as the distribution shape, centre and spread, as one cohesive whole (Garfield, delMas, & Chance, 2007). Variability within a data set is readily revealed by graphical or pictorial descriptions of distribution. Without variability there is no need for the concept of distribution. A distribution is a means of describing variability. Shaughnessy (2007) identified six different perspectives on variability:

- as extremes or outliers;
- as change over time, e.g., a time plot as a variable changes with time;
- as the whole range, e.g., the population space;
- as the likely range, e.g., used in an analysis of samples to develop notions of spread such as standard deviation;
- as the difference from some fixed reference, e.g., from a mean; and
- as a measure of the collective amount by which a distribution differs from a reference point, e.g., when comparing two sample distributions (p. 984 – 985).

Comparison of data is an important motivation (Ben-Zvi, 2004a; Watson, 2005; Watson & Moritz, 1999) to reason about variation. Konold and Pollatsek (2004) found that of the Year 8 students who could perform the calculations, only half used the mean or median to make a comparison of data sets. In his research, Shaughnessy (2006) noted a range of student responses when comparing distributions: focus on specific value without any reference to the aggregation of the data, focus on outliers or unusual values, or preoccupation with values of high frequency such as modes. Measures of centre, such as mean and median, could be used to describe and compare distributions, but are generally not used in context by students (Konold & Pollatsek, 2004; Mokros & Russell, 1995). Shaughnessy and Ciancetta (2002) showed that students who initially thought incorrectly about a probability task were likely to change their opinion after they had seen the variability that occurred from witnessing repeated trials. Shaughnessy (2006) subsequently examined students' conceptual understanding of variability in the comparison of two distributions. The entire spectrum of additive, proportional and distributional reasoning could be demonstrated.

Data distributions are conveniently represented graphically. Schools have used graphical representations of data from primary school onwards, but the concept of variation is usually first considered at high school (Department of Education, Tasmania, 2008). Students may be able to construct specific types of graph, but in terms of statistical literacy a graph has the objective of displaying features of the data such as variation, clusters, middle, notable features, or a combination of these features. Watson (2005) noted that graphical representations should take place within the larger setting of a statistical investigation, and that studying representations provide insights into student thinking.

Data and variability are two of the big ideas of statistics and as such provide topics worthy of statistics education research. The research literature notes the relationship between additive, multiplicative, and distributional thinking, and consideration of differing data representations. A shift to higher level thinking may be promoted by unequal data sets and comparison of data sets.

2.4.6 Students' notions of chance events

Random phenomena are characterised by short-term unpredictability and long-term stability (Gal, 2005). Students' development of understanding of probability, chance, and randomness have been studied over many years (e.g., Fischbein & Gazit, 1984; Hawkins & Kapadia, 1984; Piaget & Inhelder, 1975) and continue to be a topic of considerable research interest (e.g., Abrahamson, 2009; Johnston-Wilder & Pratt, 2007).

Probability is a field of mathematics where students bring their own informal beliefs and misconceptions to the classroom (Batanero & Sanchez, 2005). These beliefs may confound learning and are highly resistant to change (Batanero & Sanchez; Dunbar, Fugelsang, & Stein 2007). The beliefs are only partially related to development because both elementary and undergraduate students find such notions challenging (Metz, 1997; Konold, 1989). Researchers (e.g., Albert, 2003, 2006) have identified three distinct perspectives of probability: classical (outcomes with calculable probabilities), frequentist (estimating probabilities through repetitions of random experiments), and subjective (a numerical measure of a person's opinion of the likelihood of an event). To be effective, learning experiences must incorporate all three perspectives. Metz called for students' notions of chance to be assessed along three dimensions of conceptual understanding, beliefs, and the cultural setting, which suggests that beliefs and the community culture are an integral part of probability education.

Causal or deterministic explanations of chance processes are commonly held misconceptions by students at all levels of education (Kahneman & Tversky, 1972; Kapadia, 2008; Konold, 1989). Causal thinking assumes that every state of affairs has a cause – an explanation – and that nothing can be attributed to chance (Batanero, Henry & Parzysz, 2005). Konold argued that the use of causal reasoning was "the most significant difference between novice and expert reasoning in chance situations" (p. 92), and while students persisted with this causal view "the better part of statistical logic and all of probability theory will elude them" (p. 92). This belief and the bias to deterministic thinking may be nurtured by school experiences, for example science classes, that emphasise causal explanations (Jones, 2005).

High school students' beliefs were examined by Batanero and Serrano (1999) complementing Green's (1991) earlier large scale study of student beliefs of random number patterns. The researchers found students' performance on simple tasks increased along with age, but with more subtle tasks involving "semi-random" sequences there were "no significant differences by age in students' ability to identify distributions" (Jones, Langrall, & Mooney, 2007, p. 919). This suggests that detecting the unpredictability and irregularity described above remains challenging. Other studies of students' longitudinal developmental change in beliefs in chance phenomena (Fischbein & Schnarch, 1997; Metz, 1998; Watson, 2006) also found increasing sophistication with age along with widespread persistent misconceptions of probability.

Students' notions of chance events continue to be an on-going topic of statistics education research. Students bring to the field their own beliefs, and to be effective, the study's pedagogical approach sought to accommodate these beliefs. Achieving sustained long-term conceptual change amongst students may be difficult.

2.4.7 Students' notions of coin and die systems

Die and coin systems were an integral part of the Tasmanian school mathematics curriculum (Department of Education Tasmania, 2008). The principal value of die and coin systems lies in making a random process visible: they are physical or virtual models of a theoretical random process. The systems offered four key features to both education and education research: (a) a simple system familiar to students beyond the classroom, (b) an opportunity to model physically and mathematise random processes as a basis for more formal mathematical study of probability and statistics, (c) a recognition that many origins of probability theory lie in games of chance (Batanero, Henry, & Parzysz, 2005), and (d) a pedagogical approach that parallels the historical development of probability theory (Greer & Mukhopadhyay, 2005). Students are conceivably likely to see an electronic simulation as a simulation of a physical coin or dice, rather than as an underlying abstract mathematical process.

Watson and Moritz (2003) conducted a longitudinal study of elementary and middle school students' development of beliefs regarding a physical die. This study traced students' development of understanding of the interplay between chance and the fairness of three potentially unfair physical dice by examining both their beliefs and the strategies that the students would use to assess the fairness of the dice. Of the 108 students who participated in the Watson and Moritz study, 34 were of the same Year 9 level used in this study.

The students in the Watson and Moritz study were presented with three physical dice that were either fair or unfair. Of the 34 Year 9 students, 62% (n=21) provided the response that assumed the dice were theoretically fair, but without consideration of testing or trialling, and 17.6% (n=6) gave the most sophisticated response, where students considered both the physical characteristics of the dice and proposed systematic empirical trials. When examined on strategies to assess whether the dice were fair 62% of the students provided responses where the physical characteristics (symmetry, pip pattern, weighting) of the dice were considered and unsystematic trials were suggested, and 17.6% of students provided more sophisticated responses that considered the physical attributes of the dice, or the relationship between short-run outcomes and expected outcomes.

Their research led Watson and Moritz to develop a four-tiered system of beliefs of the fairness of dice based on the SOLO model (Biggs & Collis, 1982). This hierarchy is presented in ascending order of sophistication: (a) the die as unfair, for idiosyncratic reasons; (b) the die as fair, but without experimental justification; (c) the die as fair, with consideration of physical characteristics of the die or rolling conditions used; and (d) the die as fair in the long-run, but recognition of short-term variation. Watson and Moritz speculated that the relatively low level of sophistication of the responses reflected the fact that students had not sufficiently addressed the task or that the students were the product of a classroom culture that did not expect beliefs to be tested. Only moderate development occurred over the previous three years of the study, with half of students responding at a higher level and half unchanged in their beliefs. The persistence of erroneous beliefs was noted by other researchers (Batanero & Sanchez, 2005; Batanero & Diaz, 2007).

More complex dice systems, specifically the compound event of summing two dice, were studied by Pratt (2000), Abrahamson (2006), Abrahamson and Cendak (2006), and Watson and Kelly (2009), and are considered here because the more complex task provides further insights into students' thinking. Both the Pratt and the Abrahamson studies were small in scale and conducted with middle-school students using a series of classroom activities of either a combination of physical and virtual simulations or physical simulations alone, and the Watson and Kelly study reported on longitudinal development in a large scale study of Year 3 to Year 10 students. Pratt thought that students' naïve understanding affirmed the equiprobability bias identified by Lecoutre (Lecoutre 1992, cited in Pratt): each die was fair, and the outcome of each individual die was equally likely so the outcome of the sum of the two dice was equally likely. Abrahamson thought simultaneous presentation of different (e.g., tabular, iconic, and graphical) representations of the data supported development of deep understanding of the concepts. The Watson and Kelly study suggested only modest development of understanding from Year 6 to Year 9 and a persistence of misconceptions amongst approximately 20% of the students.

Pratt identified four local and three global resources that students use for “stochastic sense-making” (Pratt, 2000, p. 607). Local resources are constructed from short-term (short-run) behaviours of a random process, and global resources are developed through exposure to the concepts of the data aggregate. The local resources were unpredictability, where the next outcome cannot be predicted; irregularity, where no patterned sequence is identified; unsteerability, where events cannot be controlled; and fairness, related to the physical appearance of the device. The three global resources identified were: the proportion of outcomes for each possibility is predictable; the proportion for each outcome will stabilise as an increasing sample size is considered (i.e., a re-statement of the Law of Large Numbers); and the proportions can be controlled by manipulation of the sample space distribution. Pratt noted that the students brought to the study local resources based on symmetry and their experiences of random processes. Students built upon, and extended their existing local resources by adopting, constructing, and modifying global resources as part of a sense-making process. Development of understanding was considered as a “tuning towards expertise” (diSessa, 1993, cited in Pratt, p. 624) where local resources are re-structured and connected with new global resources. Multiple resources are held concurrently, “new resources do not generally replace prior resources” (Pratt, p. 624), and a distinction between “in school” and often contradictory and irrational “out of school” beliefs has been noted (Batanero & Diaz, 2007). Pivotal to the study were the activities where students re-constructed their beliefs of fairness. The technology and the activities played a formative role in developing new resources that “shifted students’ attention (away) from behaviour at a surface level” (Pratt, p. 623).

The research literature related to students’ notion of chance events guided the design of the study presented here. Researchers concluded that long-term sustained student development was difficult and transfer to unfamiliar situations was problematic because existing long-established local resources were likely to dominate students’ thinking. They also concluded that two mathematically connected and analogous situations may appear to students as distinct and unrelated and more recently acquired global resources are likely to remain dormant and are not automatically re-cued. In contrast to other researchers (Konold, 1989; Shaughnessy, 2003) who recommended directly addressing students’ misconceptions, Pratt and Noss (2010) suggested teaching methods that allowed global resources to out-compete long-established local resources by revealing the greater explanatory power of the global resources. The concurrent holding and displacement of differing beliefs has been noted by other researchers, where "to correct misconceptions it is useful for students to make predictions" (Flores, 2006, p. 291). Similarly, Watson and Moritz (2003) advocated providing activities where "questioning the fairness of the random generator be placed high on the curriculum agenda" (p. 302), and that students need experience with dice in both complex situations and in the simple context of a single die Watson (2002),. Developing confidence in empirical techniques provides the foundation for use of simulation that may be applied elsewhere. In this study Fathom is used to overcome some of these distracting cues.

2.4.8 Students' notions of sample size

Professional statisticians and mathematics educators have observed the widespread misunderstanding of sample size within the media (Harding, 1992; Kmietowic, 1994; Utts, 2003), amongst the general public (e.g., Fielding, 1996; Simon, 1997), and students (e.g., Smith, 2004; Watson, 2006). Surveys, such as electoral polls, are quoted in the media but may be misleading and misinterpreted by both journalists and the public, and the surveys may not explicitly state the sample size or consider the practical significance of the results. Consultant statisticians observed that those unfamiliar with sampling theory were often preoccupied with sample size as a fraction of the population rather than the absolute sample size (Fielding, 1996), and were frustrated when determining an appropriate sample size, saying "don't give me a complicated method just give me a rough number" (Simon, 1997, p. 389). Sample sizes are often fixed simply and irrationally by the time and resources available, or by convention. Statistics teachers and statistics education researchers (Gal, 2002; Smith 2004; Shaughnessy & Chance, 2005) considered sample size as both an element of statistical literacy, and a topic not well understood in the classroom.

Research with primary and high school students (e.g., Abrahamson, 2006; Konold, Harradine, & Kazak, 2007; Shaughnessy & Ciancetta, 2002; Stohl & Tarr, 2002) focussed only on building intuitions of sample size concepts such as that the variability of a statistic decreases with increasing sample size or making connections between classical and frequentist probability, but none of the studies considered students explicitly determining sample size. Watson's (2006) extensive work with primary and middle school students examining statistical literacy – a group that include Year 9 students similar to those in the study – considered sample size, but the work focused on part/whole concepts and, along with other researchers (Jacobs, 1999; Lavigne & Lajoie, 2007), on sample representativeness and sampling strategies. Pratt (2000), in working with young (10-11 year old) children, found that they had a natural preference for small over large sample sizes. The misconception of a belief in small samples suggests that few students possess a sound intuitive understanding of sample size, or have any formal path into quantifying sample size. Other research (Abrahamson, 2009) found, when using virtual simulations where the time required to collect data was inconsequential, that students preferred large sample sizes. This study sought to exploit the data collection efficiency that virtual simulation and Fathom provide.

Smith (2004) observed that college students, when studying sampling from large populations, found three key concepts counter-intuitive: sample size is based on an absolute number, not a proportion of the population; a larger population does not require a larger sample; and the survey accuracy depends on the actual sample size not the proportion of a population. Smith found many of her students would use a sample size rule based on the proportion of the population, such as "10% of the population." Smith formally examined college students' naïve understanding of sample size and how that understanding may develop through effective instruction. In a pre-study test, and in response to the statement: "You need to obtain a sample that is at least 10% of the population in order to get a reliable estimate of the population parameter," 76.6% of students responded incorrectly either "neutral/not sure" or "this seemed right" (p. 9). In response to the statement "For large population sizes, the size of the population is irrelevant to the reliability of the sample estimate; what matters is the absolute size of the sample" only 13.3% of the students responded correctly (p. 10). Bill, Henderson, and Penman (2010) provided evidence of similar misconceptions amongst Year 9 school students that included the belief that a sample size should be a proportion of the population and that the sample proportion declined with increasing population size, and Watson (2006) gave one example of sample size based on "10 or 20%" (of the population)" (p. 34). Smith developed a teaching program that emphasised virtual simulation, practical activities, and classroom discussion that demonstrated considerable development of understanding of sample size amongst the students in her study. This had two implications for the design of the study presented in this thesis. The first implication was a belief amongst students in the use of a sample size of 10% of the population; if students at college level had this misconception then students at Year 9 may have even less sophisticated notions of sample size. The second implication was that older college students were potentially amenable to developing more sophisticated understanding of sample size, but it was uncertain whether these notions could be developed amongst younger high school students.

High school students' first formal exposure to sample size may be activities designed to develop an intuitive understanding of the Law of Large Numbers. The law states that in repeated independent trials with the same probability of success in each trial, the percentage of successes is increasingly likely to be close to the expected chance of success as the number of trials increases (Stark, 2010). Commonly modelled using a physical coin the topic was an established part of the school curriculum used at the time of the study (e.g., Department of Education, Tasmania, 2008). By middle school students expect that the proportion of heads in repeated tossing of a fair coin will approach one half in the long run (Flores, 2006). More generally students develop an appreciation of the importance of large sample sizes when making inferences (Pratt, 2000; Stohl & Tarr, 2002). The converse – a misconception – that the law of large numbers applies to small samples, and that a small sample is highly representative of a large population was described by Tversky and Kahneman (1971) as the "belief in the law of small numbers." (p. 106). This belief, which may have an origin in the law of large numbers, leads to an underestimate of the size of confidence intervals, over-confidence in results and trends, overestimate of significance, and a tendency to attribute deviation to causal explanation rather than random chance (Sotos, Vanhoof, Noortgate, & Onghena, 2007; Tversky & Kahneman). In the context familiar to school students this belief is manifested as the misconception that within a long coin toss series short-run proportion of heads will be closer to the half than can be expected by chance; and in the roll of a physical die the frequency that each face appears is more consistent than is expected by chance (Flores, 2006) – in short, students expect the "evening out" to occur within smaller sample sizes than can be legitimately expected by chance. Flores developed a series of calculator-based classroom activities designed to build intuitions of sample size and to address the misconception of small sample sizes, but did not report findings of students' response to these activities. The misconception of a belief again suggests that few students possess a sound intuitive understanding of sample size, or have any formal insights into quantifying sample size.

Students encounter large populations through the media, for example political opinion polls, and the school orientated *CensusAtSchool* program (Australian Bureau of Statistics, 2009) now provides data sets where analysis is feasible only through sampling of the data. The literature does not formally define a large population in the school context but the populations in the pre- and post-test items used in this study presented here (e.g., Section 3.3.2.4) are arguably large. A sampling strategy of “10% of the population” described above is clearly impracticable in the large populations used in election polls. These populations are large, but they are also finite; students do encounter infinite populations at school in die and coin systems, and a sampling strategy of “10% of the population” clearly has no application in an infinite population. In this instance the sample size is chosen without sound mathematical basis (e.g., Simon, 1997), but on the time available or the endurance of students to roll a die or flip a coin.

The literature suggests the need for the development of students’ conceptual understanding of sample size and for a legitimate and accessible sample size model because: misconceptions regarding sample size exist in schools and in the wider community, sample size is largely ignored in the existing school curriculum, and the sample size used for the infinite populations of coin and die systems and the large data-sets used in schools are chosen without a sound mathematical basis. Students will have an intuitive sense of sample size and accuracy: the survey will be perfectly accurate if the entire population is surveyed. Sample size models of students include idiosyncratically or arbitrarily chosen sizes, a sample size based on a proportion of the population such as 10%, and beliefs in inappropriately small sample sizes. Alternative sample size models are addressed subsequently after first considering measurement and measurement error.

2.4.9 Measurement and measurement error

Measurement is an integral part of the school mathematics curriculum (e.g., Department of Education, Tasmania, 2008) where students count objects, or, for example, measure the familiar physical properties of mass, length, and time. Consideration of measurement error is a sophisticated extension of measurement, and in ordinary life, measurement accuracy is often largely ignored, with consideration given only that the measurement is sufficiently accurate for the purpose at hand.

Taylor (1982) considered error analysis in the physical sciences extensively. In measurement the term error does not have the natural language connotation of mistake or blunder, but it is the inevitable uncertainty associated with measurement. Measurement and its associated error may be given by the expression: Best Estimate ± Uncertainty (Taylor, p. 6), where the uncertainty was formerly described as a combination of systematic and random error (e.g., Joint Committee for Guides in Metrology, 2008). This expression illustrates three features of measurement: The measurement is an only an estimate, uncertainty is an intrinsic and inseparable part of physical measurement, and the true measurement cannot be known with perfect precision. Simple measurement of, for example, mass can be known only probabilistically: the precise mass of an object cannot be known with certainty, only that the true mass lies within a certain range and with a degree of confidence A physical measurement is more properly a sample taken from an infinitely large population of measurements where the true underlying value is unknown and unknowable because regardless of the sample size used the measurement cannot be known with certainty because a census of an infinite population is clearly impossible. In natural language an estimate is an educated guess, or a crude measurement, but in metrology (the science of measurement) what is conventionally considered measurement is an estimate only. The uncertainty may be made extremely small, but never eliminated entirely, and the only reasonable expectation "is to ensure that errors are as small as reasonably possible, and to have some reliable estimate of how large they are." (Taylor, p. 3). Minimising measurement error has an inherent cost, and such a cost must be considered against the extra benefit of any higher accuracy. The issue becomes one of practical importance and relating the magnitude of the uncertainty to the practical consequences of that uncertainty.

2.4.10 Sampling as measurement

If measurement of familiar physical parameters can be reconceptualised as sampling, then the reverse is true also: sampling can be reconceptualised as measurement. Measurement, statistics, and sampling are inextricably linked: Measurement is sampling, and sampling is measurement. A statistic derived from a survey is also a measurement – when sampling from a finite population, problems and uncertainties arise both in identifying and collecting valid and appropriate measures, and in the accuracy of measurement of the measure themselves. This study concerned itself principally with the latter: the measurement error associated with sampling.

When a survey of a population is conducted the result is a measurement of a chosen population parameter with a certain level of accuracy. Only when a census is conducted is the underlying population known with certainty, but in common with more familiar measurement the accuracy obtained from a sample may be sufficient for the task at hand. The true underlying value will be known only if the entire population is surveyed through a census, so the outcome of a survey can also only be a sample estimate of the underlying population parameter. The error associated with that sampling measurement is the standard error, which is a measure of sample variability (Shaughnessy & Chance, 2005). The standard error is used to calculate a range of values – a confidence interval – likely to contain the true or underlying population value. Traditionally, this interval is expressed as either two or three standard deviations above and below the sample estimate.

The link between conventional physical measurement and sampling was used recently by education researchers as a basis to introduce fundamental notions of variability to middle school students. Lehrer and co-workers (Lehrer, Kim, & Schauble, 2007; Lehrer, Konold, & Kim, 2006; Lehrer & Schauble, 2002), introduced the concept of variability as the precision of repeated measurement of physical objects familiar to students such the height of a flag-pole and a person's head circumference. Konold and Kazak (2008), working with students in Years 7 and 8, extended the application of exploratory data analysis to a study of probability. A preliminary activity of repeated measurement of a familiar physical object introduced students to variability in measurement, and three subsequent activities examined measurement in a virtual probabilistic environment.

2.4.11 Mathematical modelling of sample size in the classroom

Models and modelling are defined broadly in the literature and include familiar representations of graphs and tables and internal psychological representations (English, 2010). Zawojewski, 2010, cited in English, provides a definition of modelling as "a system of interest [….] represented by a mathematical system – which will simplify some things, delete others, maintain some features, and distort other aspects" (p. 26). Such a representation is clearly imperfect, but nevertheless useful. Within the study presented in this thesis modelling was restricted to formal mathematical algebraic modelling to distinguish it from other representations and mathematical techniques such as simulation processes.

The potential benefit of mathematical modelling to students' mathematical development at all levels of education is well established in mathematics education research (e.g., English, 2010). Modelling is a means of promoting higher-order mathematical thinking (English; Zbiek & Conner, 2006), enhancing the mathematical experience and skill of learners (Stillman, Brown, & Galbraith, 2010; Zbiek & Conner) by providing an additional problem-solving technique (Lesh & Zawojewski, 2006), solving multi-disciplinary tasks (Perrenet & Adan, 2010), encouraging students to focus on the structural aspects of a concept (Lesh & Zawojewski, 2006), developing relational and higher-order thinking (Chinnappan, 2010; Lesh, Lester, & Hjalmarson, 2003), and connecting with other mathematical content and mathematising tasks (Lesh & Zawojewski; Yoon, Dreyfus, & Thomas, 2010). Modelling is also an element of current curriculum documents (e.g., Department of Education, Tasmania, 2008).

Of the potential benefits of modelling identified two are particularly significant for this study, and both are designed to encourage development of deep understanding of the mathematical concepts. The first extends the development of informal intuitions of concepts recommended by statistics education researchers to developing formal mathematical knowledge through a process of mathematising, and the second is connecting to other mathematical content. Mathematising and connecting to other mathematical content are incorporated into the study in direct response to the importance of connections to other mathematics content to developing mathematical meaning (Gal, 2004; Ma, 1999 ; Noss & Hoyles, 1996); and of the belief that "a genuine knowledge of probability can be achieved only though the study of some formal probability theory" (Batanero & Diaz, 2007, p. 124).

Stillman, Galbraith, Brown, and Edwards (2007) proposed a five-step framework to implement modelling in the classroom successfully that translated a "messy world-problem" (p. 691) to an algebraic model. Stillman, Brown, and Galbraith, (2010) identified two broad approaches to modelling in education, either using a practical contextual problem to teach mathematical content, or teaching mathematical modelling purposefully using students' existing mathematical knowledge. With the exception of Chinnappan (2010), contemporary research in this topic in schools emphasises teaching the process of mathematical modelling purposefully through “model-eliciting activities” (Lesh & Zawojewski, 2006) where students develop models collaboratively. The disadvantage of this approach is the emphasis on students applying, rather than extending, their formal knowledge by the introduction of new mathematical concepts. In this study modelling was used explicitly to extend mathematical content because the sample size model used is a mathematical function not presently part of the high school curriculum.

Section 2.4.8 noted that naïve notions of sample size may be categorised into one of four broad approaches of idiosyncratic beliefs of sample size (e.g., Watson, 2006), a sample size chosen arbitrarily, a belief in small samples (e.g., Tversky & Kahneman, 1971), and a sample size based upon a proportion – for example 10% – of a population (e.g., Bill et al,, 2010; Smith, 2004; Watson, 2006). All four approaches may be imagined as models or schemes of sample size, and all four demonstrate an entanglement of beliefs, intuitions, and misconceptions. Of the four only one – the 10% of the population model – attempts to provide a mathematical model

A sample size model of 10% of a population is simply a model, and in common with all models it has limitations and simplifications. These limitations fall principally into the three broad and overlapping categories of formal statistical, pedagogical, and practical. From a formal statistical perspective the 10% of the population model, relative to more formal sampling models, under-estimates the appropriate sample size for populations smaller than approximately 800, and over-estimates the sample size required for populations larger than 800 (Appendix H.2). The model does not promote sensible and sophisticated survey design (Smith, 2004), and it does not allow direct consideration of the accuracy of a survey. From a pedagogical perspective the model does not provide a learning pathway to the formal statistical methods considered presently only at tertiary level, and it does not cultivate sense-making of the result of a survey and of measurement. Furthermore it reinforces the common misconception that the proportion of the population is an important factor in sample size selection (Smith, 2004), obscures the Paretto effect (the law of diminishing returns) that for any additional sample size increment the added information produced tends to be smaller (Watson, 2006), and it encourages less rational alternative approaches to sampling based on convention or students’ endurance or commitment to collect data. From a practical perspective the sampling model is unfeasible for very large populations encountered in national election polling, and it cannot be used with infinite populations that are encountered at school, such as die and coin systems. An appreciation of these limitations is potentially accessible to the high school students through consideration of the time and cost of conducting a survey, as a natural extension of the Law of Large Numbers activities, and by the use of electronic simulation.

This study notes the benefits of mathematical modelling to students' mathematical development. It seeks to extend students' existing informal notions of sample size and their earlier studies of the part-whole nature of a sample, as well as the ideas of random and representative sampling, to a formal mathematical model that quantifies sample size and relates the error of measurement to the sample size used.

2.4.12 Three criteria for an alternative large population sample size model

Sample size models are not part of the high school curriculum presently (e.g., Department of Education, Tasmania, 2008), or given in school textbooks (e.g., Nolan, Phillips, Watson, Denney, & Stambulic, 2000). Three criteria were used to identify and select a sample size model for this study.

1. The model must potentially lie within students' grasp. The model is accessible to the students, the underlying principles are capable of being conceptually and intuitively understood by the students, and the model can be generalised and transferred (Lesh & Harel, 2003) to contextual tasks. These criteria are included in response to criticism that statistical techniques are not applied sensibly by students, to the importance of using "known and practised knowledge and techniques" (Galbraith, Stillman, Brown, & Edwards, 2005) when teaching modelling in schools, and to contemporary statistics education that emphasises the importance of contextual tasks in developing students understanding (e.g., Watson, 2006). If the "10% of the population" sampling model is to be displaced outside of the confines of the mathematics classroom the alternative sample size model must be capable of convenient recall and application.
2. The model reveals, not conceals, key underlying concepts (Zbiek & Conner, 2006). The model should provide insights into, and connect with, fundamental statistical concepts and mathematics (Batanero &

Diaz, 2007; Chinnappan, 2010), and give "mathematically productive outcomes" (Galbraith et al., 2005, p. 5). Of principal importance in this study is that both the sample size and measurement error can be estimated formally.

3. The model recognises that students' development of statistical concepts lies on a continuum (Department of Education, Tasmania, 2008). This continuum provides a learning trajectory that builds on foundations developed in earlier years and supports an increasingly sophisticated appreciation of sample size (Watson, 2006) that allows for more formal study of statistics at senior levels. The model must be appropriate for the students' stage of mathematical development. A model may be a simplification of more formal analysis, but any simplification must not grossly contradict formal models that students may encounter at senior years.

2.4.13 The large population sample size model used in this study

The sample size model proposed for this study estimates the margin of error of a sample proportion when sampling from a very large dichotomous population, such as public opinion polling prior to a national election with a choice between two major political parties or candidates, or a choice between supporting and opposing an aspect of public policy. The model does not calculate sample size directly, but relates sample size to the error associated with the sample size used.

More formally the model relates sampling variability or the survey error, *e*, to the sample size, *n.* This margin of error provides a confidence interval within which 95% of proportions generated by randomly selecting a sample from a population will fall. The model is given in the literature (Franklin et al., 2007; Shaughnessy & Chance, 2005), it is derived in Appendix H.1, and it is presented as

$$e = \pm \frac{1}{\sqrt{n}} \quad \text{(Equation 1).}$$

The model has four assumptions: (a) a binomial experiment with an outcome classified as either success or failure; (b) samples drawn from the population randomly; (c) independent samples, that is, a constant probability of success and failure; and (d) a large sample size given as np & $nq > 5$, where p and q are the probability of success and failure respectively (Walpole & Myers, 1978; Shaughnessy & Chance, 2005).

The model also introduces four approximations. The first and second approximations are that the binomial distribution may be approximated by the normal distribution at sufficiently large sample size, and that the sample proportion provides an estimate of the population proportion. The third approximation relates to the confidence interval used. Statisticians traditionally use a confidence interval within which 95% of all randomly samples occur, which is equivalent to 1.96 standard deviations. The sample model used here reverses this logic by setting the confidence interval at 2 standard deviations first, which is the interval within which 95.4% of sample proportions will occur. The fourth approximation is that the proportion of success, *p,* is fixed and approximately equal to a half. This has the consequence that the model becomes progressively less accurate as the proportion departs from p=0.5, but the practical consequence of the error becomes less significant the further the proportion departs from 0.5 because the survey result becomes more conclusive. The four approximations simplify the model, and two of the approximations make the model conservative by over-estimating the error for any given sample size.

This sample size model addresses the three criteria given in Section 2.4.12. The model is potentially accessible to high school students familiar with the reciprocal and square root operations; it formalises the key relationship between sample size and sampling error, and it allows exploration of the function $1/\sqrt{n}$ that is used extensively in statistics. The model provides a developmental path for students from naïve notions of sample size such as “10% of the population rule” to more sophisticated sample size models studied at the tertiary level. The model can be transferred and applied to large dichotomous populations, such as political opinion polls encountered in the media.

This sample size model has been explored recently with 17–19 year old senior high school students using Fathom statistical software (Maxara & Biehler, 2010; Biehler & Prommel, 2010) in the contexts of a virtual coin toss and a quantitative refinement of the Law of Large Numbers, as well as a “rule of thumb” for sample size in simulations. Biehler and Prommel reported that on the post-study test 55.8 % of the students used the model to calculate a 95% confidence interval correctly. This suggests that aspects of this model may be comprehensible to the students in this study. The chapter now turns to graphical representations of data and data distributions.

2.4.14 Graphical data representations

Graphical representations as a means of displaying data aggregates play a widespread role in society through the media and advertising, are central to science, mathematics and statistics, and are an established part of the school mathematics curriculum. Graphical representations have the potential to support and organise thinking (Kidman & Nason, 2000) by providing an overview of the data, revealing the underlying data structure, highlighting specific characteristics and features (Bakker, 2004; Spence & Lewandowsky (1990), cited in Friel, Curcio, & Bright, 2001), and identifying relationships between variables (Friel et al., 2001). In this study graphs are used to display data distributions, and students' abilities to interpret the graphs were thought critical to their development of understanding of the mathematical concepts. A comprehensive review of this extensive research field is beyond the scope of this thesis, but published reviews of graphing research literature from a psychological research perspective (Shah & Hoeffner, 2002) and from mathematics education research (Friel et al., 2001) provide an overview. These reviews are used to define graph sense, consider the process of comprehension, and identify four critical factors influencing graph comprehension that might be applied to students' learning in this study. These reviews were largely set in traditional paper based formats; research on the use of graphs in the computer environment used in this study is more limited.

2.4.14.1 A definition of graph sense

The terms graph comprehension and graph sense are not defined robustly in the literature, and are used seemingly interchangeably. Graph comprehension was defined as "graph readers' abilities to derive meaning from graphs created by others or themselves" (Friel et al., 2001, p. 132), but this definition does not distinguish between direct reading of explicit information presented in a graph, and the more complex and abstract processes involved in inference, evaluation, synthesis and extrapolation that might be required to interpret the same graph fully. Friel et al. extended the notion of number and symbol sense to define graph sense as "a set of behaviours and ways of thinking" (p. 145). This includes the ability to recognise the components within a graph, to understand the relationships and conventions within the graph, to speak the language of the graph, to assess the information within the graph objectively, and to select the most useful form for the graph. This study uses the term graph sense, and the term graph comprehension is applied to the processes involved in comprehending a graph.

2.4.14.2 The process of graph comprehension

The process of graph comprehension has been researched extensively from perspectives as diverse as education (Ainley, 2000; Curcio, 1987; Watson & Moritz, 1999); anthropological tool use (Meira, 1998); symbolism, psychology, cognitive science, and information processing (Carpenter & Shah, 1998; Simkin & Hastie, 1987; Trickett & Trafton, 2006); business and management (Jarvenpaa & Dickson, 1988, cited in Friel et al., 2001, p. 125), mass communication and graphic design (Feliciano, Powers, & Keral, 1963; Kosslyn, 1994). More recently research has focussed on computer visualisation associated with the development of Graphical User Interfaces, computer-based imagery, and dynamic displays (e.g., *International Journal of Human – Computer Studies*). Research evidence specific to statistics and probability education in classroom-based computer environments (e.g., Abrahamson, 2006; Bakker & Gravemeijer, 2004; Ben Zvi, 2004a) examined graph comprehension, but as a subsidiary activity within other mathematical tasks, such as exploratory data analysis or data simulation.

This previous research has enabled researchers to propose models of graph comprehension based on knowledge use, and the spatial, perceptual and cognitive processes involved in graph comprehension (e.g., Carpenter & Shah, 1998; Peebles & Cheng, 2003; Trickett & Trafton, 2006). Many of the models describe comprehension as a series of steps. Bertin 1993, as cited in Carpenter and Shah, 1998 describes the process of graph comprehension as drawing on a series of three elements: translation, to interpret a graph verbally; integration and interpretation of two or more features of the graph; and extrapolation and interpolation beyond the understanding of the essence of the graph to identify inferences and consequences of the information. From the psychology research perspective Ratwani, Tafton, and Boehm-Davis (2008) examined graphing tasks using a choropleth graph (where colour coding and shading represents magnitude), and described the complex nature of graph comprehension as a combination of interactive perceptual and cognitive processes involving three stages of pattern recognition and visual decoding, identification of conceptual relationships between the features, and relating the graph referents of axes and scale to the visual features within the graph. Such studies have been criticised as narrowly focussed, laboratory-based, and lacking the social context of the education learning environment (e.g., Freedman & Shah, 2002).

Researchers have also criticised these approaches as not truly reflecting the non-linear and iterative nature of learning (Bakker, 2004; Carpenter & Shah, 1998; Konold & Higgins, 2003; Peebles & Cheng, 2003). Graph comprehension is described as a process where the graph reader shifts attention from one aspect of the graph, to another, and back again, in a process that serves to reinforce the information in memory, and mentally construct, and assemble progressively, the component "chunks of features" (Liu & Wickens, 1992, cited in Ratwani et al., 2008) into a cohesive structure. A graph is a sign: something visible that stands for something invisible or abstract. The dual nature of visibility and invisibility is described using an analogy of a window through which the outside world is revealed, but where one remains relatively unaware of the window's presence (Ainley, 2000; Meira, 1998). Bakker (2004) considers a graph as a diagram that describes a complex relationship between symbols. Symbols are developed for a specific purpose and are capable of refinement through a process of evolution and development variously described as a "cascade of inscriptions" (La Tour, 1987, cited in Roth & McGinn, 1998), where multiple translations of the information are performed until the representation reaches its final form. A graph is also an artefact – a term that is used subsequently as an element of instrumental genesis – that provides access to meaning and significance beyond the artefact (e.g., Ainley, 2000).

Graphing is a process of data translation, reduction, and aggregation. Data translation refers to the situation where data that may have once had number values now have their values defined by positions on a graph, and positions relative to other data points. Data reduction is the process where data are aggregated or recalculated, which has the consequence that the original data become invisible. Konold and Higgins (2003) noted the distinction between non-aggregated data, where each data point retained a one-to-one correspondence with the original data, and aggregated data, where a direct reference to the original data points is lost. Aggregation, whether calculating a statistic, or creating a graph, or a summary table, potentially promotes understanding by revealing the underlying structure of the data, but it also increases the level of abstraction and the risk that the meaning will be lost. To minimise the risk that meaning is lost Bakker (2004) argued that students should, at least for initial tasks, be able to trace the source of the original data, and Abrahamson (2006) and Watson (2006) sought to reinforce the link between the original data and the graph for middle-school students by using iconic representations on the graph.

2.4.14.3 Four key factors influencing graph comprehension

Having examined the process of graph comprehension the review presented here turns to consider the factors influencing graph comprehension. The two reviews of the research literature of Friel et al. 2001 and of Shah and Hoeffner (2002) complemented by other sources and more recently published literature (e.g., Bakker, 2004) identify factors that may be consolidated into four categories: (a) the tools, concepts and conventions of graphs, (b) graphs and task complexity, (c) context and familiarity, and (d) characteristics of the user. These four factors are discussed in turn.

The tools, concepts, and conventions essential for graph comprehension (Friel et al., 2001) include orientation (e.g., axes often increase to the right and "up"), format (e.g., axes on the one graph invariably have different meanings), and scale (e.g., scales may be nominal, frequency, ratio, percentage, or a combination). Scales were identified by both Rangecroft (1994, cited in Friel et al.) and Fry (1981) as a source of difficulty for students; for example, when interpolating students often misread scale where only alternate tick marks were numbered. The scale used may also affect the perceived shape of the data distribution – a problem potentially exacerbated by dynamic software graphical features that allow easy re-scaling (Ben-Zvi, 2000). Wickens (1992, cited in Carpenter & Shah, 1998) proposed the proximity principle: "best features are ones that represent the data most explicitly" (p. 97).

Particular types of graphs are inherently more complex, abstract, and less easily understood than others. Cobb, McClain, and Gravemeijer (2003) implied that value-bar graphs should be introduced before dot plots, whereas Konold and Higgins (2003) suggested introducing dot plots before histograms by applying the principle that the less the data are abstracted, the more easily the data will be understood. Large scale studies of graph comprehension (delMas, Garfield, & Ooms, 2005) of high school and college students revealed common difficulties in reading histograms and graph scales. Watson and Fitzallen (2010) in a review of the research literature argued that to promote learning it was important that "the constituent elements be made explicit and reinforce how the elements of the graph are linked to make a meaningful representation" (p. 65). The clearest guidelines for students' use of graph are provided by curriculum developers, the NCTM (2000) and the Department of Education Tasmania (2008): both propose sequences in which particular types of graphs should be introduced, and at what stage of students' cognitive and mathematical development.

Graesser, Swamer, Baggett, and Sell (1996) note that question-asking plays a central role in both cognition and comprehension, with different levels of questioning provoking – and presumably revealing – different levels of comprehension. This is described on a spectrum from low level questioning that addresses explicit material only, to high level questions involving higher order cognition of inference, synthesis and evaluation. The literature makes a distinction between tasks involving reading explicitly portrayed data, and tasks where the data must be purposefully extracted using higher-order processes of information extraction and interpretation. Curcio (1987) proposed a three-level hierarchy of task complexity: "reading the data," involving literal and explicit information; "reading between the data," involving integrating and interpreting the information; and "reading beyond the data," requiring use of the skills of inference, comparison between data sets, and prediction. Curcio found high school students had little difficulty in reading the data to answer explicit questions, produced errors associated with mathematical knowledge, interpreting language, scale and axes when reading between the data, and, perhaps predictably, found the high levels of graph comprehension required for reading beyond the data tasks of comparison and inference, prediction, and generalisation the most challenging. The three-tiered hierarchy developed by Curcio is used in the design of graphing tasks in this study, because it is consistent with GAISE guidelines (Franklin et al. 2007), and the hierarchy provides both learning opportunities for students at all levels of mathematical development and is used to support the analysis of students' responses.

Task context and familiarity involve at least two aspects: familiarity with the context and data, and familiarity with the graph format. Context provides one of the referents used to interpret the graph – a critical consideration in the study – but context is also potentially confounding because it introduces the effect of readers' prejudicial personal beliefs to the interpretation of the graph. Researchers in statistics education (e.g., Pfannkuch & Wild, 2004; Watson, 2006), advocated the use of context as a means of promoting understanding. Contrasting studies (Roth & McGinn, 1998), using a context outside a person's knowledge, also demonstrated the importance of context where users were less successful in interpreting the graph. Follettie (1980) argued that studies that used the same context with multiple representations led to one representational form being favoured over another, as some representations convey specific features of the data more effectively than others (e.g., a table of numerical data is useful for conveying precise values). Hollands and Spence (1992) and Vessey (1991) saw matching display type and task as one of cognitive fit that led to consistent problem-solving processes. Other researchers recommended promoting sense-making by utilising technology to have access to a range of data representations that allow students to choose the representation most meaningful for them. MacDonald-Ross (1977) sought to explain levels of competence in terms of exposure and familiarity with the graph format and opportunities for practice, a view consistent with Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang (2008) who noted the importance of providing students with opportunities to understand and practice key sub-skills.

The characteristics and skills of the graph reader user are highly influential for comprehension: "individual differences in graphic knowledge [....] play as large a role in the comprehension process as does variation in the properties of the graph itself" (Carpenter & Shah, 1998, p. 97). In a study of 7th, 9th, and 11th Year students Berg and Phillips (1994) found a significant positive relationship between graph ability and logical thinking and proportional reasoning; and Curcio (1987) and Gal (2002) noted the importance of mathematical knowledge and an understanding of core number concepts such as percentage and ratio in applying statistical analysis techniques. Students' interest and motivation to interpret graphs are not addressed specifically in the literature, but more widely elsewhere in statistics education research with the use of engaging, contextual, and experiential tasks. In a learning environment that cultivates discussion and dialogue, students' ability and willingness to articulate their responses to activities involving graphs is an integral part of the task.

2.4.14.4 Summary of graphs

Graph sense is a set of behaviours and ways of thinking. It includes the ability to recognise the components, relationships and conventions within a graph in a way that allows the information to be assessed objectively (Friel et al., 2001). Graph comprehension is a complex, iterative, and multi-stage process of pattern recognition and de-coding as a graph reader assembles the information into a cohesive structure (Carpenter & Shah, 1998; Ratwani et al., 2008). Graphing is a process of data translation, reduction and aggregation that may reveal the underlying structure of the data, but it also increases the level of abstraction and the risk that meaning may be lost (Konold & Higgins, 2003). Four key factors are identified as influencing graph comprehension: tools, concepts and conventions; task complexity; context and familiarity; and characteristics of the user. Principles that support comprehension are "the less the data are abstracted, the more easily they will be understood" (Konold & Higgins, 2003), "the best (graph) features represent the data most explicitly" (Wickens (1992), cited in Carpenter & Shah); task complexity lies within students' zone of proximal development (Krause, Bochner, & Duchesne, 2003); and tasks are familiar, contextual or experiential (Pfannkuch & Wild, 2004).

2.4.15 Section summary and implications for this study

Statistics education research recommends engagement with the big ideas of statistics – in this study samples, sampling and distribution – and the development of a culture of enquiry and statistical processes, active learning, data analysis using authentic data, interpretation and discourse, effective use of computer technology's visualisation and exploration features, and authentic assessment that genuinely assesses understanding. The development of intuitive notions, statistical thinking and statistical habits of mind are as important as acquiring statistical skills and knowledge (Resnick (1998), cited in Ben-Zvi, 2004a), but must be taught purposefully (Chance, 2002). Informal statistical language can be used with greater precision as statistical sense develops (Bakker & Gravemeijer, 2004). Encouraging students to consider aggregate properties of centre and variability and to move between individual cases and the data distributions may promote higher level multiplicative and distributional thinking. Informal, and often erroneous, beliefs of chance and random behaviour are strongly held, difficult to change, and may impede learning (Batanero & Sanchez, 2005).

Misconceptions of an appropriate sample size are widespread and include a belief in small samples, a sample as a proportion of the population such as a sample size 10% of the population (Smith, 2004). Formally quantifying sample size is not a present part of the school curriculum and, as such, represents a relatively unexplored topic of statistics education research in schools.

Sampling is reconceptualised as measurement with its own quantifiable uncertainty. Three criteria were used to identify a suitable sample model that it is accessible to students and it is appropriate for their stage of development, reveals key concepts, provides a learning trajectory for students, and can be generalised. The sample size model used in this study $e = \pm 1/\sqrt{n}$ relates the survey error, e, to the sample size used, n. In the study presented here this model served the dual purpose of providing an opportunity for students to examine sample size and mathematical modelling.

Graph comprehension is a complex non-linear iterative process where data are translated, reduced and aggregated. A principle that supported comprehension was that the least complex and least abstracted the data the more likely the information will be understood. Graphs were used extensively in the study presented here. The principles of graph interpretation identified in the research literature contributed to the design of the study's pedagogy, and Curcio's (1987) three-tiered structure of tasks of reading the data, reading between the data, and reading beyond the data is used in designing tasks and to support analysis of students' responses.

2.5 Computer technology and educational statistics software

2.5.1 Introduction

This section provides a brief historical perspective of the use of computer technology in schools; the software features considered desirable by statisticians, education researchers, and teachers; a comparison of purpose built mini-tools and software such as Fathom, and the role of software in the probability and statistics classroom. The theoretical framework of instrumental genesis, supported by situated abstraction, principles from education researchers introducing technology in schools, is examined with a view to effective implementation of computer software in the classroom and providing a framework to support analysis of students' responses. Computer-based virtual simulation is considered from the perspectives of both professional statisticians and education researchers.

2.5.2 An historical perspective

In 2000 Ben-Zvi noted that personal computers had only been introduced in schools relatively recently. Since Ben-Zvi's observation was published computing capacity, and the range and power of educational software, has increased significantly, and "computers, software, and the Internet are now essential tools for instruction in statistics" (Friel, 2008, p. 280). Researchers' initial enthusiasm was tempered by the realisation that the availability of the technology alone was insufficient (e.g., Ben-Zvi) and whilst the potential of the technology in mathematics education was widely recognised (e.g., Forgasz, 2006; Pratt & Noss, 2002), this potential was not yet realised (Jones, Langrall, & Mooney, 2007; Thomas & Chinnappan, 2008). The new technology could not simply be incorporated into the existing pedagogy and curriculum, and researchers and education authorities were obliged to re-assess the curriculum and the technology, how the technology should be integrated into the class environment, and how the technology should be used effectively.

2.5.3 Desirable features in educational probability and statistical software

Kaput (1992) – in a review of computer technology in education described many years later as comprehensive, challenging and seminal (Hoyles & Noss, 2008) – identified features of computers as "sources of learning efficiencies" (p. 533) that supported education. These features included:

- dynamic, rather than static, media that demonstrate continuous transition and variation – a feature particularly valuable when considering variables;
- interactive, rather than inert, systems and manipulatives that provide built-in guidance constraints and support;
- dynamically linked representations that support the cultivation of connections between numeric, algebraic, and graphical representations;
- structured and facilitated access to stored information including data sets, additional information and explanation provided "on demand," and mathematical content;
- shared computational and representational power that allows off-loading routine processes (to which Hoyles and Noss added the caveat of visible); and
- facilities that focus attention on concepts, abstractions, mathematics content, and models.

Nickerson (1995) identified features of software that promoted understanding of mathematics. Given that statistics is a mathematical science the same principles could be applied to the teaching of statistics. The software should:

- start where the students are, not the instructor;

- promote learning as a constructive process – in the sense that it should allow a student to build on existing and developing knowledge;
- use models and representations that are familiar to the novice; and
- use simulations that draw the students' attention to a situation and problem.

Kidman and Nason (2000) synthesised research to develop principles that would promote learning. The visual representations should:

- be displayed clearly and be understood explicitly;
- enable the student to focus on the deep structural aspects;
- portray physical environments;
- have external memory to display information temporarily during problem solving (This reduces the cognitive load of the person analysing the information. An analogy is solving a complex calculation on paper.);
- allow exploration and construction of understanding;
- allow translation between mathematical and natural language; and
- provide opportunities for interpretation and expression.

After identifying the deficiencies in the software available at the time, Biehler (1997) advocated a workspace with a multiple window environment, linkages for experimentation and simulation, and an analysis capability for summary results. (Snir, Smith, and Grosslight (1995), cited in delMas, Garfield, & Chance, 1999) recommended software that allowed students to perceive phenomena not observable under normal conditions (normal conditions arguably including small sample size available in physical simulation), allowed mapping between different representations, and highlighted the interplay among verbal, pictorial and conceptual representations. Ben-Zvi (2000) believed software should be coordinated, adaptable, extensible and simple. Simplicity should be the key: complexity and power should not be a barrier to the use of the software. The software should have well-defined interfaces, adequate data transfer so that familiar data can be used and dynamic object linkages between software representations and components. It should allow an interactive environment with data exploration and visualisation in a variety of forms. Software should facilitate working with data, looking at data from different perspectives, and should permit asking questions and conducting a dialogue with the data. More recently Hoyles and Noss (2008) have proposed four principles guiding the implementation of Kaput’s (1992) earlier proposals: (a) attend to representational infrastructure, so that learnable systems are available, (b) work for instructional change, so that the software takes students where and how they want to go, (c) attend to the implications of outsourcing of computational power to the computer, so that an understanding of the underlying calculations is not lost, and (d) exploit connectivity, so as to promote meaningful exchange of information.

Professional statistical analysis packages, such as MS-Excel, were not considered suitable for introductory statistics courses for children (Bakker, 2004; Ben-Zvi, 2000). Professional statistical software is too complex, and the software is designed for statistical analysis, not education; it has a high cognitive entry cost, and it does not allow the tool and the user to co-evolve (Biehler, 1997).

This study speculates on the features that software should offer practising professional teachers and their students. From teachers’ perspectives introducing and developing familiarity with the software consumes valuable class time, and for this investment in time to be justified the tool must have a low entry cost (students’ ability to use it productively in a short time period), a long-term application (students may use the tool in subsequent school years), and have a broad application (the software has application in other areas of mathematics and other curriculum areas). From students’ perspectives their willingness to use the software may be its ease of use (“user-friendliness”) and the relative attributes of the software compared with alternative electronic resources.

Significant researchers such as Kaput (1992) and others (Nickerson, 1995; Kidman & Nason, 2000) supported by statistics education researchers (e.g., Ben-Zvi, 2000; Biehler, 1997) and the researcher’s perceptions of the needs of practising teachers and school students have identified key software features that guided the choice of the software used in this study.

2.5.4 Computer mini-tools and Fathom

Bakker (2002) makes the distinction between two broad categories of statistics educational software: route-type and landscape-type. This is a distinction similar to specialised and general tools made by Olds, Schwartz, and Willie (1998, cited in Kaput, 1992), or black-box and white-box software (Buchberger, 1989), or black-box, glass-box and open-box software (Hosein, Aczel, Clow, & Richardson, 2008). Route-type, specialised, or black-box software, all have in common that the operation of the software is concealed and cannot be modified readily by the user. In contrast for landscape-type, general and white-box software the internal operation of the software is, to a degree, visible and the user has a level of control over how the software operates. The two broad categories of software have both advantages and disadvantages when learning statistics and are now discussed in turn.

Route-type software, occasionally known as mini-tools or applelets, is software specifically designed for a specific task or technique or for the development of a specific skill. This focussed approach allows the teacher to set clearly defined boundaries for the work-space, which is an advantage if classroom management and discipline is a significant issue. The disadvantage is that it constrains the students to examine only a limited number of often tightly controlled alternatives, which is at odds with current research and thinking on statistics education. Cobb (1999) used route-type software with features not generally available with commercially available software tools, but which were based on current statistical educational research. In particular, the software fitted the thinking of the user, rather than the user fitting the software. It also fitted the student’s learning trajectory by providing a familiar entry point but continued to support the student as more sophisticated statistical thinking emerged. Cobb and McClain (2004) selected software mini-tools using two criteria. They argued that software should be developmentally appropriate, which would ensure that the cognitive load and the entry time to develop a basic proficiency were relatively short, and the software should also support more sophisticated thinking as students’ thinking developed. They noted that the inherent small scale of mini-tools may also limit the scope for students’ development.

Landscape-type software is characterised by an open construction environment where students are not guided, or constrained, to take a particular route to a solution, and this allows students to follow their own individual learning trajectory. Students are also able to create their own, often unconventional, representations of data. This structure also gives teachers greater flexibility to develop their own learning sequences. Route-type software has provided software developers with the foundations, insights, and prototypes to develop landscape-type products such as Fathom. The major disadvantage with landscape tools, according to Bakker (2002), is that students are given too many options, the learning environment is too complex and students may become confused. The main objective of the lesson can be lost, and there is the potential for off-task behaviour.

Fathom (Key Curriculum Press, 2005) has been favourably received by the education research community (Hammerman & Rubin, 2004; Lane-Getaz, 2006; Lock, 2002; Maxara & Biehler, 2006). Fathom has simulation features and offers multiple ways of presenting data. It encourages students to manipulate data and transform information. The software allows samples and sampling distributions – the subject of this study – to be examined for structure, shape, and other characteristics that single statistics can only present numerically. The graphical representations encourage the development of an intuitive sense of the shape of the data distribution, and a view of the data aggregate. New technology tools, such as Fathom, do not change the complexity or the quantity of data, but give people more options for presenting information and this may aid interpretation, and ultimately, decision making. Novel presentations of data are a catalyst or boundary object (Hoyles et al., 2004) for classroom discussion. Students need to be comfortable with letting complex ideas simmer (Hammerman & Rubin, 2004) and to learn that there is rarely one clear way to make a decision when dealing with complexity and variability.

The most significant, and on-going, research into the use of Fathom in schools is by Rolf Biehler and co-workers in Germany. Having identified the need for statistics education software (Biehler, 1997) Biehler selected Fathom because, as he put it, it had tools for exploring data, tools for elementary simulations, tools for studying mathematical functions, and at the same time served as a meta-tool and meta-medium where teachers and learners could adapt working and learning environments (Biehler, 2003). Working with 17-19 year old students and with undergraduate teachers – students at least three years older than the students in this study – Biehler examined exploratory data analysis tasks (Biehler, 2003, 2006), computer based simulation of statistics and probability (Biehler, 2006; Maxara & Biehler, 2007), and more recently computer based modelling of sample size (Biehler & Prommel, 2010). The research was conducted in classroom environments with follow-up interviews with selected participants to provide more detailed information using instructional guidebooks and modifiable Fathom worksheets. Task complexity ranged from simple introductory tasks utilising data graphical display features to more complex tasks that included formal theoretical probability calculations. This research allowed the researchers to identify and categorise students' working styles (Maxara & Biehler) and further refine the teaching resources used. The most recent work (e.g., Biehler & Prommel) suggests that Fathom has become progressively institutionalised in the German senior high school system. The studies now include a preliminary fifteen lesson pre-course designed to improve the efficiency by which the software is learnt through building intuitions, learning the basic steps of simulation, and acquiring basic and stable Fathom competencies. Given the age and maturity of the research cohort the level of sophistication of formal mathematics is considerably higher than that expected of a high school student. The use of Fathom based simulation techniques is discussed subsequently (Section 2.5.7).

Lane-Getaz's (2006) progressive integration of Fathom into a senior high school statistics program provided an example of a teaching professional's use of the software. In the second year of the program, two of five topics, bi-variate data and inference, were delivered in Fathom. In the third year of the study, Fathom was used throughout the course including in the final assessment research project. Lane-Getaz concluded that students, as part of this course, demonstrated improved statistical thinking, used statistics more appropriately and accurately, and their interpretations and conclusions showed measurable improvement. The improved performance of the students was attributed to a number of contributing factors that included Fathom, the use of investigative projects, process orientated software, engaging activities employing the big statistical ideas, formative assessment and the teacher's ability to interweave topics into a conceptual whole.

The statistics education research literature identified two broad classifications of software as route-type or landscape-type software. Route-type software are task specific tools that provide highly guided learning experiences, but are potentially inflexible; landscape-type software offers greater flexibility and a variety of learning pathways, but has the risk that the students becomes confused. Fathom, a landscape-type software, offers many of the features identified as desirable by statistics education researchers. The software is also used in senior schools and tertiary institutions, and it is the subject of current research, in both Germany and the USA, and consequently it is the software used in the study presented here.

2.5.5 The role of computers and software in the classroom

Statistical education researchers recommend that software support learning, rather than occupying a central role. The software is invariably used as part of a classroom culture that promotes enquiry, discussion and active learning, exposes students to the big ideas of statistics, and uses authentic tasks and authentic assessment (Bakker, 2004; Ben-Zvi & Arcavi, 2001; Ben-Zvi, Garfield, & Zieffler, 2006; McClain & Cobb, 2001). Cobb (1999), for example, in a study of a Year 7 group, did not introduce software until lesson 5 in a sequence of 34, and Lane-Getaz (2006), in an extended classroom study, concluded that Fathom was only a contributing factor to improved student outcomes.

Ben-Zvi (2000) viewed computers specifically as cognitive tools that have the potential to improve learning. All of the studies presented by Ben-Zvi and Garfield (2004) involved the development of students' conceptual models and thinking that were largely independent of the type of technology used. Kaput (1994) noted the importance to learning of computer-based representations that facilitated a connection between human experience and mathematics.

Insights may also be gained from the recent introduction of other electronic technology, such as Interactive Whiteboards [IWB] (e.g., Moss, Jewitt, Levacic, Armstrong, Cardini, et al., 2007). One large-scale study (Glover & Miller, 2001) examined teachers' use and integration of IWB in 25 UK schools. Through analysis of 100 video-taped lessons the researchers identified three developmental stages: (a) supported didactic, where the IWB is used as a visual aid only; (b) interactive, where the IWB is used to stimulate response from the students; and (c) enhanced interactive, marked by teachers' thinking that "seeks to use the technology as an integrative part of teaching [...] to integrate concept and cognitive development in a way that exploits the interactive capacity of the technology" (cited in Goos, Dole, & Makar, 2007, p. 326). In this study Fathom was used in an "enhanced interactive" way to support learning.

Vale and Leder (2004), in a study of study of middle school students' views of computer-based learning in mathematics, reported that girls viewed computer-based learning less favourably than boys. Girls were less inclined to consider software relevant to their mathematics learning, with a tendency to see computer-based study as providing skills in computer use, not necessarily mathematics. Boys were more positive about computer-based learning, including that they found using computers pleasurable and that it supported their mathematics learning. Success with computers and students' interest were linked positively with high achieving girls and boys more positive about computer use. This research has implications for this study, which includes providing computer-based learning opportunities that gives a priority to mathematics learning over computer use and utilising boys' interest in computers.

The software Fathom has features thought desirable by statistics education features, must be incorporated into the pedagogy thoughtfully to be effective. The literature recommends that software should be used to support learning rather than play a central role and that mathematics learning must have a priority over computer use. The attitudes of girls and boys to computer-based learning may differ substantially.

2.5.6 Theoretical framework to introduce technology

Education researchers have turned recently to instrumental genesis (e.g., Artigue, 2000; Guin & Trouche, 1999; Kieran & Drijvers, 2006), to describe the process of acquisition of computer technology, and situated abstraction (e.g., Hershkowitz, Schwarz, & Dreyfus, 2001; Hoyles, Noss, & Kent, 2004; Maxara & Biehler, 2007; Noss & Hoyles, 1996; Pratt & Noss, 2010) to consider how students acquire the key underlying mathematical concepts that the technology is being used to promote. Instrumental genesis, supported by aspects of situated abstraction, and the affordances and constraints approach (e.g., Brown, Stillman, & Herbert, 2004; Guin & Trouche, 1999), are discussed in turn.

2.5.6.1 Instrumental genesis

Instrumental genesis is a process where skills and knowledge are applied to an artefact (tool) to produce an effective instrument. Drijvers, Kiernan, and Mariotti (2010) provide a review of theoretical frameworks for the use of computer technology to support learning. According to Drijvers et al. instrumental genesis has its origins in the psychological tools approach of Vygotsky (1978), the cognitive ergonomics of Verillon and Rabardel (1995), and the anthropological approach of Chevillard (1999, cited in Drijvers et al., 2010). It has been applied by French-based education researchers (Artigue, 2000, 2002; Guin & Trouche, 1999) principally to computer-aided learning, and more specifically to computer algebra systems (Kieran & Drijvers, 2006) in direct response to the perceived difficulties of introducing graphic calculator technology in

French schools (Verillon & Rabardel). More recently the instrumental approach has been utilised with the dynamic geometry software Geometer's Sketchpad™ and Cabri II+ (Hegedus, 2004; Laborde, 2001), the spreadsheet Microsoft Excel™ (Haspekian, 2005), on hand-held platforms such as Codebreaker (White, 2007), and with Fathom (Maxara & Biehler, 2007).

The instrumental approach is characterised by four keywords: artefact, instrument, schemes, and instrumental genesis; and three ancillary terms: instrumentation and instrumentalisation, and orchestration. In brief, the artefact is the bare tool that becomes an effective instrument only when the student develops psychological schemes to use the artefact effectively; schemes are the skills and knowledge developed through the reciprocal process of instrumental genesis where the student acts upon the artefact through instrumentalisation, and where in turn the artefact acts upon the student's thinking through instrumentation; and orchestration describes the teacher's role in coordinating the learning environment. The seven terms are now discussed in turn.

The artefact is the unappropriated tool. An artefact may be a material object, such as a musical instrument, a writing tool, a computer, or a non-material object such as language, mathematical symbols, or, the topic of this research: computer software. The artefact may be entirely meaningless to the user initially, or the user may be aware of its application, but be unable to use the artefact effectively. A potential user may be aware of the application of the artefact, for example, a violin, but not be able to use the artefact effectively and play it. The artefact exists, but it is unused and not appropriated by the user. In this research the artefact is defined as a blank Fathom worksheet where the software is open and operating, but entirely unmodified.

An instrument "consists of both the artifact (sic) and the accompanying mental schemes that the user develops to perform specific kinds of tasks" (Drijvers & Trouche, 2008, p. 367). The instrument includes schemes that are the mental construction (Lagrange, 2005), the psychological construct (Verillon & Rabardel, 1995), or the cognitive schemes (Artigue, 2002) needed to use the artefact in a meaningful way. The notion of instrument differs considerably from that used in natural language, where it is a tool of precision, or one used for measurement, or a musical instrument, but it is used here to describe a comprehensive composite entity. To extend the example of a violin, an instrument within the instrumental approach includes the material violin and bow, the musical score, and the associated musical knowledge and skills – the schemes – used by a musician in a coordinated manner to play the violin. This composite entity, and the coordination of its constituent elements, may describe the situation that students encounter when using Fathom in a classroom.

Schemes are “a more or less stable mental organisation, including both technical skills and supporting concepts for a way of using the artefact for a given set of tasks” (Verillon, as cited in Drijvers & Trouche, 2008, p. 369). Schemes require an understanding of the underlying principles beyond simple mechanical use of the artefact. The bare artefact alone cannot perform any meaningful activity and the instrument does not exist alone, but the instrument comes into existence only after the artefact and the mental schemes are combined (Trouche, 2004). The user develops schemes by appropriating pre-existing schemes, constructing new schemes, and integrating the schemes within use of the artefact.

The mathematics education research community has not reached a consensus regarding a definition of schemes (Kieran & Drijvers, 2006). Such debates include whether procedural activities (which are largely visible) should be considered part of a scheme (which are part invisible psychological processes), or should be considered a separate entity. In this study an instrument is defined consistent with Rabardel (2002 cited by Drijvers, Kiernan, & Mariotti, 2010), and as a fully-functioning Fathom worksheet where the user can use and interpret effectively the Fathom workspace.

The three keywords of artefact, instrument, and scheme, and how the user combines both the artefact and schemes to yield an instrument, are central to the instrumental approach. The relationship between the three keywords may be represented diagrammatically, as in Figure 2.1.

ARTEFACT + SCHEME = INSTRUMENT

Figure 2.1. Combining an artefact and psychological schemes to create an instrument.

This appropriation and integration of the artefact, of combining schemes and the artefact, is the process of instrumental genesis. It is defined as “a psychological process which leads to internalisation of the uses and the roles of the artefact, an organisation of the user-schemes, a personalisation and sometimes transformation of the tool [artefact]” (Hegedus, 2004, p. 1031). These mental schemes also include the skills to use the artefact in a proficient manner and knowledge of the circumstances in which it can be used (Drijvers & Trouche, 2008, p. 368) in an interplay between the technical – otherwise expressed as the practical or mechanical – aspects and the conceptual elements associated with the activity (Drijvers & Trouche; White, 2007).

Instrumental genesis has three principal characteristics: (a) reciprocal, where the user acts on the artefact and the artefact acts on the user; (b) personal, where the individual user appropriates and internalises schemes; and (c) evolutionary, where the instrument and user co-evolve, a process that can be complex and time-consuming.

Instrumentalisation, that a user acts upon the artefact – defined above as the blank Fathom workspace – is self-evident; however, the user does not simply use the software, but is actively involved in shaping and constructing the workspace in the same manner that a violin is first tuned and bowed.

Instrumentation, where "the subject is shaped by actions with artefact" (Hoyles et al., 2004, p. 313), is the reciprocal process of instrumentalisation. Instrumentation can be defined as a process "directed towards the subject leading to the development or appropriation of schemes of instrumented action which progressively takes shape as techniques that permit an effective response to given tasks" (Artigue, 2002, p. 250). A more concise definition of instrumentation provided by "where the artefact shapes the thinking of the user" (Drijvers & Trouche, 2008, p. 369) is used because the principal interest in this research is how students' thinking is affected; the research is not, for example, concerned with students' development of fine motor skills.

Instrumental genesis appears to be highly personal: only an individual can either appropriate or develop, and ultimately internalise, schemes. The personal nature of instrumental genesis is noted by Artigue (2002) and Haspekian (2005), and by Defouad (2000, cited in Trouche 2005, p. 201) as "instrumental genesis is not the same for all students; it depends on their personal relationship with both mathematics and computer technologies." Schemes may develop partly through social interaction, but the schemes are essentially personal and individual, and importantly for education research, this process is amenable to categorisation (Kiernan & Drijvers, 2006).

The evolutionary nature of instrumental genesis is noted by Drijvers et al. (2010), and implicitly by Artigue (2002) as "loaded progressively with potentialities […] progressively takes shape" (Artigue, p. 250). Evolution also implies two characteristics: it is time-consuming (Artigue; Drijvers & Trouche, 2008) and it is unique, sequential, and non-linear for each individual (White, 2007). Instrumental genesis is also co-evolutionary in at least two respects: both procedural skills and conceptual elements co-evolve (Drijvers & Trouche). The reciprocal nature of the instrumental genesis process allows further refinement of the diagram presented above Figure 2.1, to that of Figure 2.2.

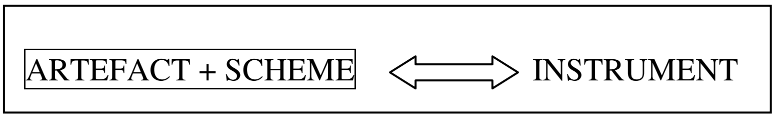


Figure 2.2. The reciprocal nature of instrumental genesis.

The process of orchestration recognises the highly social nature of learning in schools and the role played by the teacher. Orchestration, a relatively recent refinement of the instrumental approach, was developed in response to criticism (Hegedus, 2004) that the existing theory did not adequately reflect the classroom learning environment. Trouche (2004) provides a definition that is used in this research as "the external steering of students' instrumental genesis […] which includes the environmental organisation, organisation of teachers and students' work spaces and time, how the configurations are exploited" (p. 296). The metaphor of orchestration, continue Drijvers and Trouche (2008), expresses the idea of articulation and "fine-tuning of a set of instruments, guidance by the conductor, improvisation of solo players, and adaptability for different styles of music" (p. 386). The schemes developed are individual, but schemes develop through social interaction (Kieran & Drijvers, 2006). The social and environmental aspect of the instrumental genesis process allows further refinement of the representation to that presented in Figure 2.3.

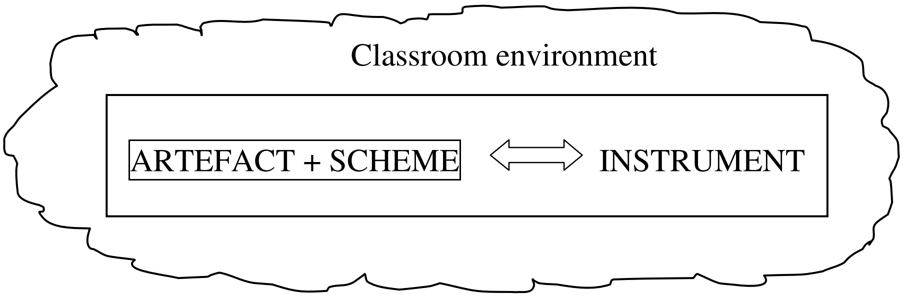


Figure 2.3. Instrumental genesis within a classroom environment.

Assude (2007) extended the instrumental approach and proposed the term instrumental integration to describe the process whereby a teacher organises the conditions for instrumental genesis, A hierarchy of five different modes of technology integration are proposed reflecting students' use of the tool, the task, and the problem-solving techniques, teacher's interaction with the task, and links to traditional techniques of pen-and-paper. In ascending order of integration the four levels of an ascending hierarchy are described, but only the lowest two levels are relevant to this study: (a) instrumental initiation where the principal aim is for students to learn how to use the technology; and (b) instrumental exploration where students explore some mathematical task. The instrumental approach may also be used for "conceptualising tool-learner interaction" (Hoyles & Noss, 2009, p. 132) that will support the learning trajectory of the use of Fathom and provide instances of tool/user interaction that promote or hinder learning, but this approach differs little from the instrument orchestration described above so orchestration is used in this study.

2.5.6.2 Situated abstraction

Situated abstraction seeks to describe how students abstract mathematical knowledge from the learning environment. Situated abstraction has been used for analysis of computer-based learning of dynamic geometry systems, algebra systems (Tabach, Arcavi, & Hershkowitz, 2008), of probability simulations (Noss & Hoyles, 1996; Forster & Taylor, 2000; Pratt & Noss, 2002, 2010), and of how learning occurred within a group learning environment (Hershkowitz, Hadas, Dreyfus, & Schwarz, 2007).

Classically, abstraction is a process of decontextualisation. It is "the act or process of separating in thought, of considering a thing (concept) independently of its associations" (Oxford English Dictionary, 1989), or a process of extracting generalisable key principles. Abstraction may be variously imagined as a shifting from the concrete to the abstract, assembling existing ideas into more complex ideas, and moving from an undeveloped abstraction to an elaborate abstraction. It is formally defined as how students "construct mathematical ideas by drawing on the webbing of a particular setting" (Noss & Hoyles, 1996, as cited in Hoyles et al., p. 321). "Situated" places students in a learning environment with a peculiar activity, context, and culture (Brown, Collins, & Duguid, 1989), "abstraction" is where students abstract the underlying mathematical principles from the situation (Pratt & Noss, 2010) or "vertically reorganizing previously constructed mathematics into a new mathematical structure" (Hershkowitz, Schwarz, & Dreyfus, 2001, p. 202), and "webbing" describes a complex structure that supports learning. Developing students' ability to abstract underlying mathematical concepts and make connections to other mathematics is reported widely to be unsatisfactory and superficial (e.g., Hollingsworth, Lokan, & McCrae, 2003; Stacey, 2003).

Mathematics education researchers have recently broadened the definition to consider abstraction as one of extraction, reorganisation, and construction: Herhkowitz et al. identified abstraction as a three-step epistemic process of constructing new knowledge, recognising an existing mathematical structure is relevant, and building-with or applying knowledge to a new context (p. 196). To reflect the non-linear nature of the elements the three elements are linked as a nested, rather than linear, structure. Abstraction is also highly iterative, and the influential role of learners' history and the learning environment is emphasised. This definition embraces the development of mathematical understanding without necessarily the development of a comprehensive understanding of a mathematical concept. The three element model of analysis of student actions during abstraction was used by Hershkowitz et al. (2001) to analyse high school students' learning of mathematics, and the model was used in this study to support examination of students' development of understanding of mathematical concepts.

Webbing is an extension of the more common term scaffolding, and both terms are designed to convey a sense of a structure that supports learning. Scaffolding is the assistance of an expert or adult or teacher that provides the appropriate level of support for students to extend their knowledge and skills (Krause, Bochner, & Duchesne, 2003). More recently researchers (e.g., Pea, 2004) have noted how the definition of scaffolding has evolved, including to the use of technology. Noss and Hoyles (1996) extended the metaphor of scaffolding to introduce webbing, which describes the process where "the student infers meaning by coordinating the structure of the learning system, including the knowledge to be learned, the learning resources available, prior student knowledge and experience, and constructing their own scaffolds by interaction and feedback" (Noss & Hoyles, as cited in Hoyles et al., 2004, p. 319). Webbing is a more comprehensive term than scaffolding because it embraces students' own background, the learning environment, and students' own active contribution to learning – a concept that more truly reflects the complex classroom environment with its multitude of factors influencing learning. In scaffolding this support is withdrawn progressively as the student develops skills and knowledge to work independently, but in webbing this support evolves as the student uses and reconstructs the structure. In this study webbing is used in preference to scaffolding.

Pratt and Noss (2010) proposed design principles to promote abstraction that include allowing students to test personal conjectures, selecting a task seen by students as generating purposeful activity, using greater explanatory power to displace existing fragmented knowledge, linking purpose and utility to promote key mathematical concepts and blurring of the informal and formal to provide a natural connection to mathematical concepts.

Situated abstraction complements instrumental genesis because it embraces aspects not specifically addressed by the latter (Hoyles & Noss, 2008) by emphasising the development of understanding of the underlying mathematical concepts. It does, however, shares at least three significant characteristics with instrumental genesis: it is time-consuming, complex, and social (Kieran & Drijvers, 2006). Firstly, time is measured on a relatively long timescale of "years rather than days or months" (Hoyles, et al., 2004, p. 322). Secondly, it is a complex process (Artigue, 2000), driven by an individual's diversity of webs and prior knowledge, where learning takes different trajectories and does not occur in a strict sequence of steps. Thirdly, situated abstraction is highly social (Trouche, 2005) with learning occurring within a community of dialogue and discussion, and it is a mutual act. The importance of discussion and dialogue within situated abstraction is recognised through the notion of boundary object, which provides a point of common focus where mathematical meaning develops through mutual construction, interaction and feedback.

According to Drijvers et al. (2010) the significant contribution of Hoyles and Noss is how the notions of webbing and abstraction are applied to understand more fully learning in a computer aided environment and how this learning can make connections with other mathematical settings. Webbing and situated abstraction are important contributions to an understanding of student learning, but the approach is not unique to a computer learning environment, and the concepts do not directly attend to the role of technology, which is a major focus of this study. Situated abstraction complements the instrumental genesis framework because it draws attention to the key objective of development of the underlying mathematical concepts involved, and it could be argued that the situated abstraction should be readily absorbed into the instrumental approach as an integral feature of instrumental genesis (Hoyles et al., 2004). In this study the language of situated abstraction – Hershkowitz et al.'s three-step epistemic process of constructing new knowledge, recognising an existing mathematical structure is relevant, and building-with or applying knowledge to a new context – is incorporated into the instrumental genesis model.

2.5.6.3 Affordances and constraints

Instrumental genesis, supported by aspects of situated abstraction, provides an over-arching philosophical framework to introduce and use technology effectively, but more specific aspects of technology use are needed to support practising professional teachers. These aspects of technology use are variously identified as affordances and constraints (Brown et al., 2004; Kennewell, 2001), or agents and constraints (Kaput, 1992), or potential and constraints (Guin & Trouche, 1999), or encouraging and inhibiting factors (Forgasz, 2006) Biehler (2006) focussed purely on the impediments to learning, which he identified as obstacles or break-down points. Put simply an affordance is the potential for action of the technology, and obstacles inhibit the potential for action (Thomas & Chinnappan, 2008). In this study the technology provides an affordance for action, and examples of constraints are the beliefs and attitudes of the students and the teacher, the time available, and the existing curriculum.

Using data collected from Australian teachers Forgasz (2006) noted the three most significant factors encouraging use of computer technology as students' enjoyment and the quality of the software, the availability of computers, and teachers' confidence and skills; the factors of developing students' mathematical skills and computers as a teaching tool rated lowly. The three principal factors that discouraged computer use were the availability of computers; teachers' confidence and the perceived need for professional development; and the time constraints of the curriculum, students acquiring basic skills, and setting up the computers. Forgasz thought that these factors differed little from similar surveys conducted in other countries and in other disciplines, and that little change had occurred during the previous decade, highlighting the difficulty of introducing technology into the classroom effectively. In this study the constraint of teacher confidence, time and computer resources, were not relevant, but the time-constraint of students acquiring basic skills and setting up the computer were still factors.

Ruthven and Hennessy (2002) surveyed school teachers to determine "what practitioners conceive as the successful use of computer tools [...] to support mathematics teaching and learning" (p. 47). This led to the development of the operational themes of enhancing ambience, alleviating constraints, assisting tinkering (i.e., students' exploration), stimulating students' motivation and engagement, facilitating routines, and supporting the development of conceptual ideas. The researchers found – highly relevant for the investigative program in this study – that "computer use was particularly important in making investigative activities accessible to students and viable in the classroom." (p. 79). More recently Ruthven (2008) incorporated affordances and constraints within instrumental genesis to characterise the acceptance of new technology in the classroom.

Biehler (2006) applied the affordances and obstacles approach in a study of Year 11–13 students using Fathom to examine data distributions. A focus of the study was how Fathom fostered or hindered learning, and whether the obstacles lay with the statistical content, the software, or the interface between the two. Biehler concluded that the data interpretation tasks should be separated from the software related tasks, and that teacher-prepared Fathom worksheets allowed students greater opportunity to focus on tasks. This section now turns to the principal feature of computer technology used in this study: computer-based simulation.

2.5.7 Computer simulation and re-sampling

Computer-based simulation in statistics education has been studied extensively over the last fifteen years (e.g., Biehler, 1994; Nickerson, 1995), and the topic continues to be of considerable interest to both education and the education research communities (e.g., Abrahamson, 2006, 2009; Biehler & Prommel, 2010; Konold & Kazak, 2008; Stohl & Tarr, 2002). This interest is driven by the curriculum frameworks (NCTM, 2000; ACARA, 2009); the widespread recognition of simulations potential as a learning tool (e.g., Forgasz, 2006; Hesterberg, 2006); the affordances provided by the increasing availability of computers in schools; the increase in processing speed, power, user-friendliness and flexibility of commercially available computer software (e.g., Fathom); and the variety of platforms available in schools including graphic calculators (e.g., Zimmerman & Jones, 2002).

A fundamental purpose of simulation is "to make the abstract concrete" (Hesterberg, 2006, p. 391). In the study presented here the abstraction being made concrete is a hypothetical random process. A computer may appear to simulate a coin-toss (or a rolling a die), but both the physical coin and the computer are more accurately simulating a hypothetical random process – one simulates the random process physically, the other virtually.

This study had a particular emphasis on one aspect of simulation: re-sampling. Re-sampling is the process of repeatedly drawing samples from a population. The term re-sampling was used in preference to the term simulation, because simulation can be applied to, for instance, video games that many students enjoy as leisure activities or simple electronic mimicry of physical processes such as coin toss. Re-sampling is an umbrella term for the formal mathematical simulation techniques of Monte Carlo, permutation tests, jack-knifing and boot-strapping (e.g., Hesterberg, 2006), but none of these terms are potentially meaningful for high school students. The term re-sampling was used in the study as a term readily comprehensible to high school students and one that accurately describes a process of repeatedly drawing samples from a population.

Recent statistics education research has focussed on the development of mathematical concepts and intuitions in an informal environment with middle-school students (Ireland & Watson, 2009; Konold & Kazak, 2008; Lehrer, Kim, & Schauble, 2007; Stohl & Hollebrands, 2006), in formal mathematical study in senior high school (Biehler, 2006; Zimmermann & Jones, 2002) and undergraduate introductory statistics courses (Chance & Rossman, 2006; delMas, Garfield, & Chance, 1999, Zieffler & Garfield, 2007), but research in high school is less extensive. In an informal statistical environment students are neither provided with, nor expected to use, formal statistical tests extensively because students, particularly those at middle-school, may not have the necessary mathematics or the number sense (Konold & Kazak). The students may, however, have preconceptions and intuitions that provide the basis for the developmental of key foundational statistical concepts that should provide the basis for a more formal approach at a subsequent stage of school. A lack of mathematical skill does not necessarily bar access to theoretical underpinnings: a fair die, a fair coin, and the sum of two die all have probability distributions accessible to middle school students (Pratt, 2000; Pratt & Noss, 2002). Two researchers (Simon, 1997; Wood, 2005) have suggested an approach to probability education entirely within simulation, and Rossman (2008) suggested a role for simulation in informal inference. In this study the high school student cohort was amenable to a partially formal mathematical approach midway between informal approaches adopted for middle-school students, but less formal than used with more senior students.

The software Probability Explorer was used by Stohl and co-workers (Stohl, Rider, & Tarr, 2004; Stohl & Tarr, 2002; Tarr, Lee, & Rider, 2006) in the culminating activity *Schoolopoly* where students investigated the fairness of a subtly biased virtual die, but none of the studies considered students' beliefs about virtual dice. Ireland and Watson (2009) extended the earlier work of Watson and Moritz (2000) with middle school students exploring sample size with a die system using both physical simulation, and, for larger sample sizes, computer simulation with TinkerPlots. Significantly for the study reported here the students expressed doubts about the simulations producing random and fair outcome. Watson (2006) noted the importance of allowing students to explore the fairness of random generators "if the full potential of a study of fairness was to be realised" (p. 170), a view echoing a much earlier recommendation of Kaput (1992).

Maxara and Biehler (2006) and Biehler and Prommel (2010) reported on the use of Fathom simulation in senior high school and in an undergraduate mathematics teachers' probability and statistics course that integrated frequentist and formal theoretical approaches. The program subsequently incorporated subjective aspects of probability, and the pedagogical model was extended to a sequence of identifying intuitive expectations; building a stochastic model and a simulation plan; analysing the task theoretically; comparing the theoretical analysis, the simulation and the intuitive theory to resolve misconceptions; and applying the techniques to new tasks. The topics and the level of mathematics required to complete the course successfully were not suitable for the high school students in this study, but the general principles presented in their pedagogical model were appropriate.

2.5.8 Section summary and implications for this study

The desirable features of computer statistics software developed by Kaput (1992) and Nickerson (1995) include that it allows multiple means of representation of the data, dynamic linkages among the representations, simulations, and convenient access to a wide-range of data resources. The software is endorsed by several researchers (e.g., Maxara & Biehler, 2006), but classroom based research with Fathom is limited. Education researchers (e.g., Ben-Zvi & Garfield, 2004) recommend using the software to support learning and to promote the development of statistical thinking and reasoning. All of the recent research, without exception, considered software as just one of many classroom resources. Fathom was chosen for this study because it satisfies many of the criteria thought important by statistics education researchers and it is used in senior schools with students several years older than those who participated in this study.

Instrumental genesis considers the introduction and effective use of an unfamiliar tool, and it was used in this study's mathematics learning environment both to guide the introduction and use of the software and subsequently to help support the interpretation of students' use of the software. Instrumental genesis (e.g., Drijvers et al., 2010) describes the process of how the artefact (the tool) is combined with users' skills and knowledge (schemes) to produce an effective instrument. Instrumental genesis is time-consuming, personal and complex, reciprocal (where the individual acts on the instrument and vice-versa), and highly social, and these principles guided the introduction and use of the Fathom software. Instrumental genesis approaches analysis from the perspective of tool-use and includes, as part of that process, schemes to use the tool generalisable skills effectively and efficiently, and to construct mathematical meaning.

Situated abstraction complements instrument theory by describing how mathematical knowledge may be developed, how students extract mathematical knowledge from the learning environment, and what kinds of mathematical knowledge are extracted (Noss & Hoyles 1996; Pratt & Noss, 2010). The three element model of Hershkowitz et al. (2001) – the language of situated abstraction – of constructing new knowledge, recognising existing mathematical structures and applying knowledge to new contexts was used to support analysis of students' responses.

2.6 Assessing students' learning

Garfield (2003) noted that in addition to pedagogy the other major area of statistical education reform was student assessment. Traditional assessment has been criticised as too narrow in not adequately describing student understanding. Authentic assessment principles (Archbald & Newmann, 1988) are used in the Teaching for Understanding framework (e.g., Darling-Hammond, Ancess, & Falk, 1998; Blythe, 1998) that has, as one objective, truly assessing the development of students' understanding. Assessment also plays a role in learning by conveying to students what knowledge and skills are of value (Garfield, 2003; Garfield & Chance, 2000; Ben-Zvi, 2004b). Student assessment using curriculum frameworks, rubrics and portfolios, and hierarchical assessment such as Structure of Observed Learning Outcomes (SOLO) models are now considered. Instrumental genesis, introduced earlier in Section 2.5.6.1 to guide the introduction and use of Fathom software in the classroom, is re-utilised to examine students' development of use of the software, and situated abstraction allows consideration of students' abstraction of mathematical meaning from the activities and tasks.

2.6.1 Curriculum frameworks

The classroom component of the study served two purposes: data collection for a research study and a teaching unit broadly compatible with the existing curriculum and assessment framework. Students in this study were two streamed extended Year 9 mathematics classes. Comments made by the colleague teachers indicated that lessons should be designed and assessed for Level 4 of the Essential Learning framework (Department of Education Tasmania, 2003). The Essential Learning Guide, Being Numerate support materials and associated documents show that a student at this level should be able to demonstrate an understanding of data to:

- show that the same relationship may be expressed in tables, words and graphs;
- justify the selection of tools (in this case within Fathom);
- perform computations with whole or part of the data; and
- represent, interpret, and draw inferences from the data.

The disadvantages of this assessment framework were that it was restricted to the Tasmanian Education system, it was recently introduced, and assessment was largely unmoderated, and consequently the curriculum framework was used to guide assessment only.

2.6.2 Rubrics and student portfolios

A second approach to assessment is the use of assessment rubrics developed for a particular statistics course. Lane-Getaz (2006) used an assessment rubric that consisted of six criteria with greatest weight given to analysis, communication and evaluation. When designing the rubric, Lane-Getaz reflected on how students could demonstrate moving beyond mechanical analyses, synthesize the available information, and be evaluative. Biehler (2003, 2006) and Maxara and Biehler (2007), in studies with Fathom software, used a range of assessment items and student portfolios, but within a traditional assessment approach that emphasised correct responses, the transfer and application of skills to related tasks, and the identification of obstacles, working styles, and commonly occurring difficulties. Rubrics and portfolios were used in this study to provide school assessment for the students, but are not presented as part of this thesis.

2.6.3 Hierarchical assessment models

Another approach to assessment is the use of stage-wise or cognitive development models that recognise a spectrum of understanding of concepts or students' development over time. Hierarchical models provide a scaffolding structure with descriptors of each level that can guide teacher assessment and support students' shifts to higher-order thinking, support instructional design, develop learning trajectories, implement instruction, and assess students (Jones, Langrall, Mooney, & Thornton, 2004). Hierarchical models also allow students' responses to be coded, classified, and analysed (Shaughnessy, 2007).

Two such development stage models are the Bloom's Taxonomy (Anderson & Krathwohl, 2001) and the Structure of Observed Learning Outcomes (SOLO) model developed by Biggs and Collis (1982) and further refined by Collis and others (e.g., Campbell, Watson, & Collis, 1992; Pegg, 2003). Both SOLO and Bloom's Taxonomy are hierarchical models of increasing complexity designed to assess higher-order thinking and cognitive learning outcomes independent of the task involved (Biggs & Collis).

No evidence of Bloom's Taxonomy applied to contemporary mathematics and statistics education was available. The SOLO model is, however, "widely used in assessing the development of students' statistical reasoning" (Jones et al., 2005, p. 99), an opinion amply supported by the extensive research literature (e.g., Pfannkuch & Wild, 2004;

Reading, 2004; Watson, 2006). The researchers (Chance et al. 2004; Mooney, 2002; Nor & Idris, 2010; Padiotis & Mikropoulos, 2010; Watson & Callingham, 2003) have used the SOLO model to construct formal hierarchical models of statistical literacy. Many of these models grew from consideration of what it is to be statistically literate, or students' development of understanding of specific concepts in statistics, such as interpretations of data. Watson and Moritz (2003) used SOLO to trace intellectual growth over an extended period of time in longitudinal studies, but it is students' responses to intensive instruction that are the subject of this study.

2.6.4 Assessing thinking using the Structure of Observed Learning Outcomes (SOLO)

The Structure of Observed Learning Outcomes (SOLO) modeldescribes development as a combination of the two features of the mode and the response, where the mode refers to the abstractness of the task and the response is the level of sophistication of the person's responses to the task (Pegg, 2003). Both the mode and the response are hierarchical.

The first key feature of SOLO is the modes of thinking, and the three modes pertinent to this study are, in order of increasing sophistication, the ikonic, the concrete-symbolic, and the formal. Individuals operating in the ikonic mode internalise actions by developing individual words and images. Individuals operating in the concrete-symbolic mode are still tied to their own concrete experiences where, for example, a few specific instances satisfy the reliability of the rule (Collis, 1975); this has implications for the study presented here where students examine the large population sample size model. Students working in the formal mode "consider more abstract concepts described as working with principles or theories where students are no longer restricted to concrete referents" (Pegg, 2003, p. 242).

Pegg (2003) noted that much of elementary and early secondary school teaching is adapted to suit students working at the concrete-symbolic level, and Panizzon, Callingham, Wright, and Pegg (2007) claimed that 14 and 15-old students in their study would work principally in the concrete-symbolic mode and display some elements in the formal mode. The modes do not preclude an individual working in a lower mode in any context, and individuals who respond in one mode may be unable or unwilling to respond in the same mode in a different context. An individual has the option of operating at a lower level; the formal mode, for example, does not necessarily subsume operations in the concrete-symbolic mode (Pegg). All of the modes continue to develop throughout life. The three modes are presented in summary along with indicative age ranges in which the thinking first occurs (Pegg, p. 242).

1. Ikonic (from two years) where the action is internalised in the form of images and language. In adults, this leads to a form of knowledge referred to as intuitive thought.
2. Concrete symbolic mode (from 6 or 7 years) is demonstrated where a person thinks through written language and symbols to describe systems that have an internal logic. Transition to this mode from ikonic represents a major increase in abstraction.
3. Formal mode (from 15 or 16 years) is where individuals seek to understand the relationship between concepts. This represents a further increase in abstraction and a reduced reliance on concrete referents.

The second key feature of SOLO is the level of response, which is an individual's ability to respond with increasing sophistication to the task. This study used SOLO's five-tiered framework to assess students' responses:

1. prestructural (P) responses where the task is not understood, or where irrelevant aspects provide a distraction;
2. unistructural (U) responses where only one aspect of the task is presented;
3. multistructural (M) responses where several disjoint relevant aspects are processed but not integrated;
4. relational (R) responses demonstrating an integrated understanding with coherent structure and meaning; and
5. extended abstract (EA) responses, which go beyond relational to include information from outside the task.

An example illustrates the use of SOLO. When applied to an exploratory data analysis task examining marathon race times, a unistructural approach would be a single statistic such as "…the fastest time occurred in 2002…" A multistructural approach might be demonstrated by a graph of race-times against year, showing a progressive reduction in times. A relational analysis would provide an integrated and appropriate response selectively and appropriately using all the information within the task which might include the change in race-times against year, the significance of gender, and the performance of specific athletes and country of origin.. An extended abstract response might explain a progressive reduction in race times on improved training methods or diet of the athletes.

The level of response is age related and the indicative ages at which levels of higher-order thinking occur are presented in Table 2.1. The subjects of the study were 14-15 years old so relational responses were anticipated.

Table 2.1.
SOLO Model. Summary Table for the Concrete-Symbolic Mode[1]

Level	Acronym	Indicative age range (years)	Indicative characteristics in the students' responses
prestructural	P	4-6	did not understand question, did not complete the work, or where irrelevant aspects provide a distraction
unistructural	U	7-9	used single features of task
multistructural	M	10-12	used a range of evidence but not in an integrated way
relational	R	13-15	combined salient features to the task at hand, provided evidence of multiple pathways, displayed a coherent and comprehensive understanding of all elements
extended abstract	EA	16+	drew in features from outside task, showed understanding of underlying principles

[1] Adapted from Biggs and Collis (1982)

A refinement of SOLO is where learning is seen as occurring in learning cycles (Campbell et al., 1992; Panizzon et al., 2007). This refinement was introduced in response to the observation that a single cycle analysis inadequately captured students' learning (Pegg, 2003). Campbell et al. developed a two-cycle analytical framework where the relational response of the first stage became consolidated into the unistructural response of the second stage, i.e., the learning achievements of the first cycle provided a basis for learning in the second cycle. In essence the first and second cycle analysis attempts to describe the iterative nature of learning where initial incomplete understandings are subsequently assembled into a more consolidated response. The researchers, denoting the first and second cycles by subscripts, presented this framework diagrammatically,

$$U_1 \rightarrow M_1 \rightarrow R_1 = U_2 \rightarrow M_2 \rightarrow R_2 \text{ (Campbell et al., p. 296).}$$

Campbell et al. applied the framework in a longitudinal study of students' development of understanding of volume from primary through to secondary school level. The first cycle referred to students' conceptual development of understanding of volume and application of the formula for a simple rectangular prism, and the second cycle referred students' application of the formulae to composite figures or rectangular and triangular prisms. Students' responses to the classroom items presented here, where students apply the large population sample size formula to a contextual task, correspond to a second cycle analysis (Callingham, private communication May 27, 2011).In this study second cycle analysis is used for the one task tha will be known subsequently as the Mt. Wellington cable-car task.

Several education researchers have criticised SOLO. Shaughnessy (2007) thought the boundaries between the levels used in SOLO were blurred, and Chick (1998) and Chan, Tsui, Chan, and Hong (2002) noted that consistent assessment using SOLO, even amongst trained assessors, was difficult. Chan et al. also considered SOLO to be unstable, in that the one assessor may assess the same work differently at different times. Such characteristics are a significant disadvantage of SOLO. It would limit the potential to compare research studies, and it may limit the ability to compare an individual student's performance amongst tasks.

Chan thought the ambiguity of SOLO levels could be reduced by the use of sub-scales. Otherwise known as transitional responses, sub-scales were first proposed by the developers of SOLO, Biggs and Collis, in 1982. Transitional responses occur when "the student is feeling for the next level, but doesn't quite make it… [the responses] carry more information than would be expected in the level the student is emerging from but [the student] is forced to give up before reaching the complexity at the next SOLO level" (Biggs & Collis, 1982, p. 29). The initials P, U, M, R and EA (see Table 2.1) are used extensively in Chapter 4, and the subscript T is used to describe transitional responses. For example, M_TR is a transitional response between multistructural and relational levels. Such a refinement also provides an additional level of granularity in assessment of students' responses in this study.

SOLO is used extensively in the statistics education research, and although it has limitations it was the formal assessment model used in this research study.. A short-coming noted by the researcher is that although SOLO was used to assess a student's development at the conclusion of the task, it did not necessarily allow the student's learning trajectory to be identified. This perceived short-coming is addressed in the subsequent sub-section.

The study was conducted as a normal teaching unit, so formal feedback to students in the same and familiar format used by the schools was offered to the two colleague teachers. Authentic assessment principles identified in the Teaching for Understanding framework (Blythe, 1998), the Essential Learnings program (Department of Education, Tasmania, 2003), and the school's normal assessment protocols were to provide feedback.

2.6.5 Instrumental genesis, situated abstraction, and affordances and constraints

Instrumental genesis, supported by aspects of situated abstraction, which were introduced in Sections 2.5.6.1 and 2.5.6.2 respectively, provided lenses both to design and view the learning process. The frameworks do not provide a formal generalised structure to assess learning, but were used to describe and explore how learning occurred in a technological environment.

Instrumental genesis provided a means for "conceptualising tool-learner interaction" (Hoyles & Noss, 2009, p. 132) that supported the learning trajectory of the user of Fathom. Instrumental genesis approached analysis from the perspective of tool-use, and included as part of that process identifying schemes both to use the tool generalisable skills effectively and efficiently and to describe how the software supported the construction of mathematical meaning. Instrumental genesis was included in this study to provide the perspective of learning within a computer-based environment. In this study instrumental genesis was also used to chart the longitudinal development in the use of the Fathom software.

This study incorporated into instrumental genesis aspects of situated abstraction to inform students' development of understanding of the mathematical concepts. The three-step epistemic process of Hershkowitz et al. (2001) of constructing, recognising, and building-with provided the terminology to describe how students used the software and other resources to develop meaning of the concepts. Situated abstraction was used to provide insights into the development of understanding of mathematical concepts that supported the explicit determination of sample size.

The SOLO model and situated abstraction were used in this thesis to complement each other. Situated abstraction was used in response to the researcher's belief that the SOLO model did not provide an effective means to identify how mathematical meaning is constructed. The SOLO model measures learning at specific points – but it does not necessarily show students' learning pathway between points of assessment.

The affordances and constraints approach was used to identify specific instances where the pedagogy and technology either supported or hindered the learning process. These specific affordances and constraints are designed to provide teaching principles that can be used directly by practicing teaching professionals.

2.6.6 Section summary and implications for this study

The SOLO model was used in the classroom phase of the study to provide a framework to describe students' understanding of the concepts and to demonstrate the range of students' responses presented. Students' responses were assessed on a five-tiered hierarchy. Instrumental genesis provided a lens to examine students' development of tool-use and learning in a technologically rich environment, situated abstraction provided the language use and means of examining how students abstracted mathematical meaning, and the affordances and constraints approach was used to provide insights into student learning trajectories and to identify how the technology supported or hindered leaning.

2.7 Implications for research and the three research questions

Statistical literacy, the knowledge and dispositions required to evaluate statistical information critically, is an essential skill for modern life in employment, informed citizenship, and practical day-to-day living. Given the importance in modern society of data and statistics the topics are an important part of the formal school curriculum.

The research literature examined the education research theoretical frameworks, the "big-ideas" of statistics, statistics education best-practice pedagogy, statistics education software, frameworks to introduce software, and the assessment model of SOLO to assess the development of statistical thinking. Instrumental genesis provided the framework used to examine students' development of the use of Fathom, and situated abstraction provided the framework to examine the development of concepts associated with sample size. The rationale for the three research questions and the three research questions are presented in sequence.

The beliefs and misconceptions of probability and statistics that students bring to the classroom may confound learning. The study sought to develop students' acceptance of virtual simulation as an essential foundation for the subsequent study of statistics and probability at more senior school years. Such a process must be both effective in promoting acceptance of the simulation and efficient because of the limited class time available. The education research literature emphasised the importance of supporting the development of sound intuitive notions of statistics through developing a culture of statistical enquiry and classroom discussion; hence students' acceptance was cultivated through a process of scientific enquiry of the random behaviour of the simulation.

> **Research question 1: How effective is a statistics education research best-practice based approach of scientific enquiry in developing high school students' acceptance of the Fathom virtual simulator?**

Misconceptions about sample size exist widely in the community. The explicit determination of sample size has not been a topic of education research nor a part of the high school curriculum, but it is a natural complement to the representative and random sampling topics that are studied in schools. Virtual simulation provides a mechanism to explore the explicit determination of sample size and to support the development of sound intuitions of sample size. Sampling is re-conceptualised as measurement with an associated measurement error. In large populations this error may be calculated using the large population sample size model $e = \pm 1/\sqrt{n}$, which relates *e,* the margin of error, to *n,* the sample size.

> **Research question 2: In what ways does the sample size model $e = \pm 1/\sqrt{n}$ provide an accessible method for high school students to explicitly determine sample size when sampling from large and infinite populations?**

Fathom offers many of the features recommended by statistics education research. The use of Fathom in Australian schools is relatively unknown, and research into the effective use of the software is not extensive. Instrumental genesis, used elsewhere to support the introduction of computer-based tools, supported by aspects of situated abstraction provided, both a framework to inform the introduction of the software and a means of analysing students' use of it. Virtual re-sampling, one of the features of Fathom, is not used extensively in schools presently, but Fathom makes re-sampling potentially accessible to high school students.

> **Research question 3: In what ways does this study's pedagogical approach of using Fathom virtual simulation and re-sampling offer an effective learning opportunity for high school students? What affordances and constraints do students encounter?**

Chapter 3 explores the application of the education research literature to the methodology for the study's three research questions.

CHAPTER 3 Methodology

3.1 Introduction

This research study was a mixed-method, naturalistic, exploratory study conducted in two Year 9 high school classes in Hobart, Australia. The study examined the research questions that explored the three themes of students' development of acceptance of the Fathom probability simulator, the explicit determination of sample size when sampling from large populations, and Fathom re-sampling as an effective mathematics learning opportunity in high school.

This chapter is presented in eight sections of the design principles of the study, descriptions of the work samples used in the classroom teaching sequence and assessments, the post-study student questionnaire, the researcher's professional journal, the interviews with the two colleague teachers, the detailed study, the procedures used for data analysis, and the chapter concludes with the work samples re-presented and grouped by the three research questions.

3.2 The design of the study

3.2.1 Overview

The study was conducted as a three week classroom teaching unit taught by the researcher in collaboration with two colleague teachers and a detailed study of twelve students in three-hour workshops conducted approximately six weeks after the conclusion of the classroom unit. The classroom and the detailed studies were designed to complement one another: the classroom study provided the context of the conventional classroom environment and the detailed study provided the forum for a focussed examination of issues thought significant in addressing the research questions.

This section describes the four phases of the classroom study, a summary of the teaching sequence, the detailed study of six student pairs, the study time-line, the sample of participants and the research setting, the data collection instruments, the six statistical tools or concepts introduced by the researcher, the principles guiding the introduction of Fathom into the classroom, instructional and peer support, the assessment frameworks, ethical considerations, and processes to promote data validity.

3.2.2 The four over-lapping phases of the classroom study

The classroom teaching sequence was conducted in a series of four over-lapping phases. Phase 1 was a series of pre-tests designed to establish students' background knowledge of basic mathematical skills considered essential for the study and their naïve beliefs of quantifying sample size. This phase introduced students to Fathom, using an exploratory data analysis task. Phase 2 was a series of classroom activities designed to promote students' acceptance of the Fathom simulation as a mathematics tool. Phase 3 was a series of classroom activities that used re-sampling to examine the large population sample size model. Phase 4 was formal post-study student assessment. The four phases of the study are discussed in detail in Section 3.3.

3.2.3 The study presented as a summary of the teaching sequence

The four overlapping phases of the classroom study lay within a conventional teaching sequence. In principle the lesson plans presented to the two schools were identical, but in practice the lesson plans differed slightly to reflect the different teaching unit times (50 minutes and 60 minutes in the boys' and girls' schools respectively) and particular circumstances that arose such as a lesson cut short for school administrative reasons or the students' productivity and motivation affected by the time of the day or the day of the week. The lesson sequence for the boys' class is presented in Table 3.1, In the study the MS-PowerPoint presentations provided the lesson plans, but lesson plans based on the presentations were written subsequently to provide lesson overviews. The presentations and lesson plans are attached in Appendices A.19–A.39. The presentations have been edited to remove material not used in the classroom,

Table 3.1
Boys' classroom teaching sequence

Lesson No.	Phase	Principal objectives	Worksheet presentation lesson plan	/ Method / section	Appendix
0	1	Pre-test students' mathematical and statistical skills considered essential for the study.	Pre-test	3.3.2.1	A.1, A.19 & A.30
1	2	Introduce basic Fathom operations using an exploratory data analysis task. Fabricate a home-made die for subsequent testing.	New York Marathon	3.3.2.5	A.2, A.20 & A.31

Table 3.1 (cont.)

Boys' classroom teaching sequence (cont.)

Lesson No.	Phase	Principal objectives	Worksheet / presentation / lesson plan	Method section	Appendix
2	2	Test home-made die, develop a statistic to assess the fairness of the dice, display data on a poster as a boundary object for class discussion, and test the factory-made die. Informally analyse the fairness measure data, and introduce GICS. Assess students' understanding of the fairness measure in a homework item.	Home-made die worksheet, Factory-made worksheet, & Fairness measure homework.	3.3.3.1 – 3.3.3.4	A.3, A.4, A.5, A.21 & A.32
3	2	Introduce Fathom simulation by assembling and testing a Fathom die simulation, calculate fairness measure for a Fathom virtual die, determine the mean, median, and mode from a dot plot.	Fathom virtual die – first Fathom simulation.	3.3.3.5,	A.6, A.22 & A.33
4	2 & 3	Compare three dice using the GICS framework. Assess students' naïve understanding of sample size Mt. Wellington cable-car.	Compare three dice using GICS framework. Mt. Wellington cable-car (naïve)	3.3.3.6 & 3.3.4.7	A7, A13, A.23 & A.34
5	3	Assess students' development of use of Fathom using a 3-minute Fathom basic skills test. Examine the effect of sample size on the fairness measure by first re-calculating the fairness measure as a percent fairness measure and second considering sample sizes of 30, 300, and 3000 tosses of a coin. Assess students' understanding of proportion of heads measure 50 & 500 tosses of a coin using a homework item.	The effect of sample size on the %fairness measure (boys' version). Coin measures 50 & 500 tosses of a coin homework.	3.3.4.1	A.8, A.9, A.10, A.24 & A.35
6	3	Introduce coin simulation and the law of large numbers using a physical coin toss. Assemble a Fathom coin simulation and collect proportion of heads at 50 and 500 tosses of the virtual coin. Examine the effect of sample size on the measure of proportion of heads	Physical coin toss (cumulative proportion of heads). Fathom virtual 50 & 500 tosses of a coin simulation.		A.11, A.12, A.25 & A.36

Table 3.1 (cont.)

Boys' classroom teaching sequence (cont.)

Lesson No.	Phase	Principal objectives	Worksheet / presentation / lesson plan	Method section	Appendix
7	3	Introduce the large population 1/(square root sample size) sample size model using a frequentist approach and a Fathom coin simulation. Test the model at sample sizes of 50, 100, 400, 900, and 1600 tosses of a virtual coin.	Large population sample size model.		A.14, A.26 & A.37
8	3	Apply the large population sample size model to the contextual task of the Mt. Wellington cable-car public opinion survey.			A.27 & A.38
9	4	Post-study assessment of students' understanding of sample size and Fathom simulation methods.	Post-study test. Fathom basic skills test. Students' post-study questionnaire		A.15, A.16, A.18, A.28 & A.39
10	4	Follow-up testing to determine long-term retention of their development of understanding of sample size by re-presenting the national and state election item.	National and state election worksheet		A.29

3.2.4 The detailed study of six student pairs

The detailed study of three all-girl and three all-boy student pairs was conducted as a 20 minute interview held on conclusion of the classroom studies, and a three-hour interview and workshop at the University of Tasmania approximately six weeks after the classroom study. The workshop of a series of eight tasks explored students' development of use of Fathom, interpretation of numeric probability data representations, use of key terminology, and the interpretation of graphs identified in this study as measures dot plots. The methodology is presented in Section 3.7.

3.2.5 The time-line of the study

Data collection for the study was conducted during the period 2007 to 2008. The pilot study conducted in 2007 helped guide the design of the classroom study conducted in

2008, but no data from the pilot study was used directly in this thesis. A time-line of the study is presented in Table 3.2.

Table 3.2.
The Time-line of the Study

Date	Event
April – Sept. 2007	Literature survey and study design
September 2007	Informal approach to the all-boys school for pilot study
October 2007	Social Science Human Ethics Committee formally approved study (Appendix B.2)
November 2007	Tasmanian Department of Education formally approved study (Appendix B.4)
December 2007	Pilot study conducted at the all-boys school
March 2008	Formal invitation to the two schools to participate in the study
May 2008	Consent to participate in study
June – Ju1y 2008	Classroom data collection stage
September 2008	Detailed study data collection
October 2008	Colleague teacher interviews and sample size follow-up test

3.2.6 The sample of participants

Five criteria were used to select the class cohort: (a) students had sufficient background mathematical skills and knowledge to allow the sample size model to be potentially accessible; (b) time and flexibility existed within the mathematics course; (c) the class had access to a set of computers; (d) school, colleague teachers, and students were supportive; and (e) students had the potential to benefit from the opportunity.

The most junior level of school – Year 9 – at which the topic of explicitly determining sample size could be conducted, was determined by the mathematical concepts involved. The large population sample size model (Section 2.4.13) required calculating surds and the reciprocal of surds, and these topics are not considered until early high school. The most senior school level where the study could be conducted – Year 10 – was determined by the flexibility within the existing mathematics course; at senior high school Years 11 and 12 students are preparing for formal tertiary entrance examinations, so schools and teachers were reluctant to offer access for a study that did not address the existing mathematics curriculum. Students who were likely to benefit most from the opportunity presented by the study had demonstrated ability and motivation to study mathematics, such as might be demonstrated by enrolment in an elective advanced mathematics class.

The study was conducted with two Year 9 extended mathematics class at two single-gender government funded high schools in metropolitan Hobart. Although the student group was an extended mathematics class, the students had self-selected to enrol in the mathematics course, and both colleague teachers believed the group was of mixed ability. The students were 14 or 15 years old. Twenty-one male and 35 female students were enrolled in the classes. Two female students were international exchange students and did not have English as their first language, so their work samples were not included in the analysis.

The students who participated in the detailed study had also participated in the classroom study. The colleague teacher was asked to select and approach prospective participants using the general criteria that the students approached should offer a range of abilities, likely to engage in the work and produce meaningful data, be moderately motivated, willing to place their work under greater scrutiny, and be willing to use the Captivate screen capture software. The students were not representative of the classes, but the students did offer a range of perspectives. Students were paid a modest gratuity of A$20 and a movie pass in recognition of their contribution to the study. The research items used in the detailed study are presented in Section 3.7.

The colleague teachers, one female and one male, were the two senior mathematics teachers at their schools. Both were career teachers in their early fifties with thirty years professional experience, teaching in rural and metropolitan schools across the full range of student ability and interest. Most significantly, both teachers had had several years experience teaching senior mathematics at the Tasmanian college level Year 11 and 12, and this was thought important because it provided a longitudinal sense of the curriculum and allowed teachers to see the topic and concepts within the broader mathematics curriculum. The colleague teachers supported the project, and were always in attendance, observing the research study, offering suggestions, and supporting the study by participating in classroom discussion and offering support with behavioural management issues on the rare occasions issues occurred. The colleague teachers were invited to participate in the refinement of the teaching unit and training, but both declined citing other professional teaching commitments.

3.2.7 The research settings

An all-boys and an all-girls government school in Hobart participated in the study. The two schools were formally invited to participate via the school principals, who referred the invitation to the head mathematics teachers.

General indicators of each school's culture included a high level of compliance with the school's uniform policy and a low level of graffiti on school property. Enrolment in the all-girls school was eagerly sought with student entry restricted by residency in the immediate area, or by sibling or maternal relationship. The students were drawn from a middle-class socio-economic group. In 2010 on the Index of Community Socio-Educational Advantage [ICSEA] the all-boys school was rated at 984 and the all-girls school was rated at 993, which is marginally below the mean score of 1000 for Australian schools (Australian Curriculum, Assessment and Reporting Authority [ACARA], n.d.).

The physical environment is a factor in teaching and learning. The research study was conducted in the schools' computer laboratories. The researcher used Fathom as a teaching aid with a data projector that projected the image of the computer screen onto a screen at the front of the classroom. Interactive whiteboards were not available. The computer laboratory in the all-girls school was used intensively and setting-up was done in the few minutes between classes, and this frequently disrupted the start of the lesson. The equipment in the all-girls school was not reliable and this also disrupted the class. In the all-boys school the computer system was largely trouble-free.

In the classroom study students used Department of Education personal computers running Windows 97-2003 operating system and Fathom Version 2. Fathom was installed on each computer in the computer laboratory, students were assigned to individual computers, and the computers were adjacent to each other. Students were able to work independently or collaboratively if they wished.

Students accessed pre-prepared Fathom work files from the school's computer system, and students submitted work electronically using a shared drive using a filename based on their personal identity code (Section 3.2.13); this was an unfamiliar practice and students needed support initially to use the directory system. In the post-study assessment students submitted work as a hard-copy. Students' electronic work samples were copied as a class set to a flash drive and saved to a secure drive at the university.

The detailed study was conducted in an office in the School of Education, University of Tasmania. Students attended, in pairs and at a pre-arranged time, during the school term vacation some six weeks after the conclusion of the classroom study. Students sat side-by-side, which allowed students to view each other's work, in an environment akin to a small classroom. The digital audio recorder provided a record of the workshop.

3.2.8 Summary of data collection instruments

There were eight opportunities to collect data:

1. Student work samples of a pre-study basic skills check, students' beliefs of a multiple coin toss and sample size, and an introductory exploratory data analysis task using Fathom (Section 3.3.2);
2. Student classroom and homework work samples that included development of acceptance of the Fathom simulation (Section 3.3.3) and explicit determination of sample size using the large population sample size model (Section 3.3.4);
3. Students' post-study assessment of key terminology (Section 3.3.5.1), sample size (Sections 3.3.5.2 & 3.3.5.6), the sample size function (Section 3.3.5.4), measures dot plots (Sections 3.3.5.3 & 3.3.5.5) and the development of the procedural use of Fathom through a basic skills test (Section 3.3.5.7);
4. Students' post-study questionnaire (Section 3.4);
5. Researcher's professional journal, including a record of the lessons, extracts of the whole class discussion, comments from individual students and the colleague teachers' observations (Section 3.5);
6. Colleague teacher interviews (Section 3.6);
7. A detailed study of six student pairs (Section 3.7); and
8. A post-study test item examining sample size that was conducted approximately two months after the conclusion of the study (Section 3.3.5.8).

The items and tasks for the four phases are presented in separate sub-sections. Each sub-section provides a summary table of all items (Tables 3.3 – 3.6) that includes the title of the item, a summary of the objectives, and references to the location of the results and the reference to the item presented in full in the appendices. Each item and task is then described separately and in detail. A consistent format is used to present each item of the title, principal objectives, a description of the activity, and the methodology used to analyse the data. The research items used in both the classroom study teaching sequence and in the detailed study are presented in the same order as they were presented to the students. The works samples are then re-presented grouped by the three research questions (Tables 3.8 – 3.10).

Interview protocols for the detailed study and the colleague teacher interviews were developed from principles identified by Kvale (1996) and Rubin and Rubin (2005). The three authors describe an interview as an intimate and complex conversational partnership. Kvale imagines an interview study as a sequence of stages that includes the designing the interview, addressing ethical issues, posing main and follow-up questions, developing strategies to improve quality, transcribing the interview, and handling, analysing, and interpreting the data. The interview protocols for the student and colleague teacher interviews are attached as Appendices C.1 and D.1.

3.2.9 Six statistical tools or concepts introduced by the researcher

Six pedagogical tools or concepts were introduced by the researcher to support students' learning. These were the following.

(a) The Global, Individual, Measures of Centre and Measures of Spread (GICS) framework was developed by the researcher (Bill, 2007) in response to statistics education research that found middle-high school students perceive data as a collection of individual points rather than as an aggregate (e.g., Ben-Zvi, 2004b), and that students would benefit from a support or webbing structure to describe data sets. GICS provides an informal framework to promote discussion by obliging students to examine the information from four perspectives – Global, Individual, measures of Centre, and measures of Spread – as an interpretation step before drawing conclusions regarding the data distribution. The framework was introduced to students as a natural extension to students' informal analysis of a graph of a series of coin tosses displayed at the front of the class. The framework was modelled by the researcher through whole-class discussion where students contributed their observations of the data and the researcher classified the feature as either a global, an individual, a measure of centre or a measure of spread aspect of the data. The students subsequently used these cues and contributed to the discussion further by classifying the data feature themselves. This approach required students to gather all available information before analysis, it encouraged reflection about the data, it promoted a culture of enquiry and statistical habits of mind, and it provided a foundation for higher level analysis.

(b) This study re-defined the term statistic. Formal definitions of a statistic (e.g., Oxford Concise Dictionary of Mathematics, 2009, p. 430; Stark, n.d.) exist, but in this study a statistic was defined informally for the study as a number that represents a more complex set of numbers. Formal statistics, such as mean and standard deviation have become established and acquired their own names by virtue of their usefulness. Informal statistics are used widely, for example, performance ratings that allow sport-persons to be listed in order of ability. Both formal and informal statistics were used in the study.

(c) The fairness measure statistic was an informal statistic used in this study to measure the fairness of a die, computed as the sum of the differences between the observed and expected frequency. It allowed students to focus on the concept that a statistic represented the more complex situation of a die rolled several times.

(d) The Fathom statistics education software was the statistical software tool used throughout the study. Consistent with statistics education research best-practice Fathom was used to support learning only, and students were not expected to develop a high degree of fluency or competence with the software. Its introduction and use in the classroom are discussed in more detail elsewhere (Section 3.2.10).

(e) The study introduced the statistical technique of re-sampling, which is the process where a number of the same statistic (in this study the term measure was used) was collected. In conventional sampling a sample of size *n* is taken and one only measure (e.g., mean) is calculated. Re-sampling repeats sampling with the sample size *n*, and more than one measure is collected. To make the distinction between the sample size and the number of measures, the expressions "sample size used to calculate the measure" and "number of measures collected" were used. The expressions, although cumbersome, were designed to minimise students confounding the two as sample sizes.

(f) The large population sample size model $e = \pm 1/\sqrt{n}$, where n is the sample size and e is the margin of error associated with measurement, was introduced. This model provides an estimate of the error associated with sample size when sampling from a large or infinite dichotomous population. More formally the model calculates a 95% confidence interval, or in the terminology of the study the margin of error, wherein approximately 95% of all simulation results will occur. In 30 simulations used typically in class approximately one or two results would be expected to occur outside of the interval bounded by the margin of error. The model was chosen using the three criteria given in Section 2.4.12: it is potentially within students' grasp; it reveals, not conceals, key statistical concepts; and it places students' development of understanding of statistical concepts on a continuum to senior school. The model changes the focus from sample size to jointly considering sample size and the associated error. This model was designed to build on students' intuitions that a survey does not provide certainty and can only approximate the underlying population, and to extend those intuitions to a more formal mathematical approach. The large population sample size model was introduced to the students without proof or derivation, as this would require a level of mathematical skill and knowledge well beyond the most exceptionally able Year 9 student, and, as an alternative, the model was demonstrated empirically using a frequentist approach and Fathom.

3.2.10 Principles used to introduce Fathom into the classroom

The objective was not highly proficient use of the Fathom software tool, but that the software tool became a part of students' mathematical repertoire progressively and allowed the students to focus on the underlying mathematical concepts under investigation rather than dedicating intellectual effort into using the software or being distracted by the software. Students were not expected to design simulations or develop a level of proficiency that allowed students to work independently because the software was not available outside of the classroom. Students were provided with the opportunity to acquire skills sufficient for the task only. The software supported learning indirectly through promoting acceptance of the simulation and cultivating students' sense of accomplishment and self-efficacy. Students' development of use of the software was intended to be both efficient and productive.

The study adopted the principle from statistics education research of using the technology tool in an entirely new way to exploit the potential of the tool, rather than simply incorporating the software into existing professional practices or extending existing practices. An example of incorporating the software into existing practices is to substitute physical simulation with virtual simulation without attending to students' acceptance of the software tool or using statistics education best practice pedagogical principles. An example of extending existing practices is to include a limited number of features such as exploring the large sample sizes readily available in virtual simulation. The study used virtual simulation in a novel way for traditional education practice by introducing re-sampling to a high school classroom.

Instrumental genesis provided the philosophical framework to introduce and use Fathom in the classroom and subsequently to analyse students' responses. Instrumental genesis recognises that students' familiarisation and internalisation of any tool can be complex, time-consuming, reciprocal, and personal (Section 2.5.6.1).

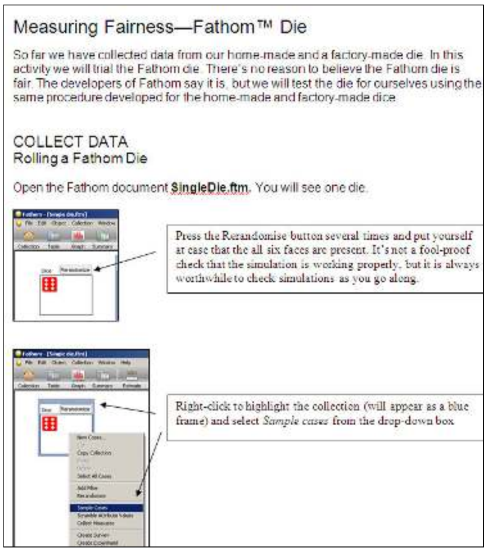
The importance of the development and refinement of statistical language as part of the process of statistical enculturation is well recognised in the statistics education literature. In this study an element of this vocabulary was the terminology used in the Fathom software, and this terminology was not necessarily familiar to students, or identical to that used elsewhere in the mathematics curriculum. In Fathom the term case is a datum point, an attribute is a data variable (e.g., height), a collection is a data set, and a measure is a statistic (Fathom, 2005).

Fathom's modular nature lends itself readily to the use of a constructivist approach. Students assembled the modules into a functioning simulation and progressively checked that the simulation behaved as anticipated. This step-wise assembly and checking approach had several objectives that included building confidence in the software and creating ownership of the simulation, providing opportunities for practice in assembling simulations and key simulation sub-skills, and slowing students to encourage reflective thinking about the simulations. Each subsequent simulation extended students' knowledge of Fathom by introducing an additional software feature. Students progressively developed a basic repertoire of skills, used key terminology, and acquired a set of procedures in the software. This established a classroom routine that could find application elsewhere with Fathom:

1. assemble the simulation and the data representation from the individual modules,
2. methodically check that the simulation behaved as anticipated, and
3. critically evaluate the data generated as a deliberate sense-making process.

The education literature categorises software as either black-box route-type software where the functionality of the software is largely set and obscure, or as a white-box landscape tool where the user constructs the simulation or where the functionality is visible (Section 2.5.5). The term grey-box was introduced in this study to describe the software's use along a spectrum between white and black-box use. Constructing the simulation was somewhat of a misnomer because student did not construct the simulation at the fundamental computer coding level, but instead assembled the simulation from its component modules, so the term assemble was used.

Students assembled the Fathom simulations guided by hardcopy worksheets (e.g., Figure 3.1). To cultivate a routine the worksheets adopted a consistent approach that had three key features of (a) screen-grabs of Fathom, (b) dialogue boxes giving specific instructions, and (c) arrows indicating the screen location of the operation, feature, function, or drop-down box. The screen-grabs allowed students to compare the appearance of their own simulation with the model provided in the worksheet, and this reassured students that the simulation was being assembled correctly.



Measuring Fairness—Fathom™ Die

So far we have collected data from our home-made and a factory-made die. In this activity we will trial the Fathom die. There's no reason to believe the Fathom die is fair. The developers of Fathom say it is, but we will test the die for ourselves using the same procedure developed for the home-made and factory-made dice

COLLECT DATA
Rolling a Fathom Die

Open the Fathom document **SingleDie.ftm**. You will see one die.

Press the Rerandomise button several times and put yourself at ease that the all six faces are present. It's not a fool-proof check that the simulation is working properly, but it is always worthwhile to check simulations as you go along.

Right-click to highlight the collection (will appear as a blue frame) and select *Sample cases* from the drop-down box

Figure 3.1. Section of an example of a Fathom worksheet.

The worksheet also provided a short series of questions designed as a natural entry into more formal analysis of the simulation and as a boundary object to promote classroom discussion. The approach adopted here was to offer a range of task complexity within any one activity to cater for the inevitable diversity of student abilities within any class group. For even the least able students the relatively straight-forward task provided an opportunity and the satisfaction of successfully assembling a simulation. The Fathom worksheets are attached in Appendices A.6, A.8, A.9, A.12, A.14, and A.16.

3.2.11 Instructional and peer support

The tasks became progressively less supported and more complex throughout the unit of work. The assessment tasks lay on a spectrum of instructional support but can be grouped into three broad categories:

- Pre-instruction or limited instruction. This was designed to determine students' background and intuitive understanding of the topic prior to tuition.
- Instructed / webbed / semi-independent / co-operative learning tasks. The environment was essentially that of a normal classroom environment combining teacher-led instruction, whole-class discussion, and student collaborative tasks. The two homework items are included in this category, because it is not certain that students worked independently.
- Independent work by students. This was a traditional examination environment that allowed an evaluation of students' learning.

3.2.12 Assessment frameworks

The Structure of Observed Learning Outcomes (SOLO) model (Section 2.6.4) was used to assess students' work samples principally because the model is used extensively in statistics education research. The use of SOLO in this study is described in more detail in Section 3.8.

Instrumental genesis, supported by the terminology of situated abstraction, was used to provide qualitative descriptive material that supported the thesis, rather than used as a formal analytical structure. Instrumental genesis was used to analyse students' use of Fathom from the perspective of tool-use, and it provided the longitudinal development framework and the learning trajectory to use Fathom effectively and efficiently. Aspects of situated abstraction were incorporated into instrumental genesis as a lens to examine students' development of mathematical meaning of the fairness measure, explicitly determining sample size, language and use of terminology, and numeric and graphical data representations.

The affordances and constraints approach (Section 2.5.6.3) was used to identify aspects where the pedagogy and technology either supported or hindered the learning process. These specific affordances and constraints were designed to provide specific teaching principles for practicing teaching professionals.

Assessment and feedback are important motivating factors for students, and formal assessment for inclusion in the students' school assessment also demonstrated to the students that the study had the full support of the school. To support learning students were provided with informal feedback during the class and formal written feedback at the subsequent lesson. Student portfolios were used to provide a more comprehensive assessment of students' understanding and to cultivate students' sense of ownership of a body of work.

In the all-girls school the colleague teacher accepted the offer for the researcher to assess the portfolios using the school's assessment protocols, with the grade included in the students' mathematics assessment after moderation (e.g., Appendix G.3). In the all-boys school the colleague teacher chose to include selected elements only in the students' formal assessment, but from the study's perspective this also conveyed to the students that the study was a legitimate part of the school's mathematics course.

3.2.13 Ethical considerations

Any novel teaching approach, including that used in this study, carries inherent risk. This risk was minimised by examining the education research literature, designing the research program within the Tasmanian curriculum framework (Department of Education, Tasmania, 2005), pre-testing the students for mathematical skills thought essential for the study (Section 3.3.2), and reviewing the unit using continual feedback from both the colleague teachers and the students participating in the study.

The study had the formal approval of the Tasmanian Social Sciences Human Research Ethics Committee (Appendix B.1), the Department of Education, Tasmania (Appendix B.2), and the consent of the schools, teachers, students and parents (Appendix B.9).

Information sheets and consent forms were provided to the Education Department of Tasmania, the principals, the colleague teachers, and to the students and the parents of students participating in the detailed study, but only information sheets were provided to parents and students in the classroom study. Participation in the detailed study was by informed consent, i.e., both students and their parents provided a signed consent form, and participation in the classroom study was by benign consent, i.e., students or their parents had to withdraw their consent actively by informing the colleague teacher. All students invited to participate in the detailed study consented, and no student declined to participate in the classroom study.

Students used a simple code to protect their identity, both to reduce any personal bias during the study and to preserve participants' anonymity for publication. The code LDDMMF (in which L=last letter family name, DDMM=birthday and F=first letter first name) was easy for the students to remember and use, and it provided sufficient information to allow the researcher to identify students if errors in the code occurred.

From the schools', the colleague teachers', and the students' perspectives the researcher acted in the role of a teacher, and in that role participants would demand at least the same standard of personal conduct as practising professional teachers. The researcher's personal conduct was guided by the protocols provided by the Teachers Registration Board, Tasmania (2006, 2007).

3.2.14 Data validity

Internal validity of the research study was maintained through triangulation and a multi-method approach that provided information from a number of perspectives: any interpretation or conclusion that is consistent from multiple perspectives has a greater level of authenticity and more complete understanding of student learning (Section 2.2.6). The use of a range of worksheet tasks also provided students with the opportunity to demonstrate a range of skills and understandings and to cater for different learning styles.

The researcher conducted the study as teaching a unit of work, so it was inevitable that the researcher assumed the role of participant observer (Tedlock, 2003). Although this allowed the researcher to have direct and rich involvement in the study, it carried the inherent risk of introducing biases and differing interpretation. Triangulation was used to reduce this risk by providing verification and consistency.

The research was conducted within the regular timetable and classroom of the students. A naturalistic approach that mimicked the normal school environment was chosen because this setting provided information on the environment in which learning occurs normally, and this may allow the results of the study to be generalised to other schools. The students were aware they were taking part in a research study, but the atmosphere of the classroom was that of a new topic with a new teacher.

3.3 Classroom teaching sequence work samples

3.3.1 Introduction and the four phases of the classroom study

The classroom teaching sequence is presented here in full and in the same order as presented to students, but not all items presented to students were used in the subsequent data analysis. The classroom sequence was grouped in four phases that were linked to the three research questions (Section 2.7). Phase 1 provided pre-study testing and background information on the students. Phase 2 and Phase 3 were teaching sequences where Phase 2 sought to cultivate students' acceptance of the Fathom simulation as legitimate, and Phase 3 provided students with the opportunity to examine sample size. Phase 4 provided the post-study assessment that examined students' development for all three research questions.

Phase 1. Establish a baseline of students' number-sense of fractions and percentages, gain an understanding of students' intuitive sense of the distribution of the proportion of heads from 50 tosses of a coin and the distribution of the frequency at which faces occur in 30 rolls of a die, and assess students' ability to analyse and interpret the dot plot graph format used in the study. Fathom was introduced to students using an exploratory data analysis task of New York marathon race times that provided a base-line of students' ability to interpret a data distribution and examined students' first use of Fathom's basic functions, modular structure, and terminology.

Phase 2. Develop acceptance of the Fathom simulator. The literature review noted that the beliefs and misconceptions of probability that students bring to the classroom are present across all stages of student development, are difficult to change, and may confound learning (e.g., Batanero & Sanchez, 2005). Statistics education researchers recommend that students are given opportunities to make predictions and challenge their beliefs, to test simulations, and to develop expertise by allowing more robust global resources and principles to out-compete existing local knowledge (e.g., Pratt, 2002). This phase of the study sought to develop students' confidence in the Fathom die simulator through an objective scientific statistical enquiry that examined the fairness of physical dice and the virtual Fathom die. The three dice examined were identified as a home-made die – a die that students fabricated themselves using Sculpey^{TM} modelling clay; a conventional factory-made die – a term chosen purposefully in preference to a "real" die that may suggests that the Fathom die is not legitimate; and the Fathom simulation virtual die. Class discussion supported students' progression from naive and informal perceptions of fairness to a formal measurement of fairness of the dice using an objective fairness measure statistic. Instrumental genesis (Section 2.5.6.1) incorporates the notion of schemes, which are the mental organisation and structure, the skills, and the supporting concepts to use the software in a meaningful way. Schemes are reciprocal (where the tool acts on the user and vice-versa), personal, and evolutionary. Such a definition does not address subjective beliefs of probability or the simulator as a legitimate mathematics tool explicitly, but this study extended the use of schemes to incorporate students' acceptance and confidence that the data generated by the simulator were legitimate – in the eyes of the students that the simulation was fair and the data generated were genuine. Much of the research of beliefs in probability is based on physical simulation models, and this study extended this earlier research to consider students' beliefs of virtual simulation. This also increased the task complexity because students considered both physical and virtual dice. Without confidence in the virtual simulation this study speculated that learning was likely to be superficial, any misconceptions were likely to persist, and further learning using the software undermined. The development of schemes, such as changing beliefs, are time-consuming processes, but such an investment is justified by purposefully examining the legitimacy of the simulation through statistical enquiry and meaningful mathematical activity that objectively examined the fairness of the simulation using the fairness measure statistic. This phase served three objectives simultaneously by attending to students' beliefs of the legitimacy of simulation, developing basic familiarity with Fathom, and modelling the process of scientific enquiry. This phase provided the classroom tuition for Research Question 1 largely in its entirety.

Phase 3. Introduce and apply the large population sample size model (Section 2.4.13) using Fathom. The principal objective of this phase was to introduce and use the sample size model in a way that potentially would convince students of the model's usefulness. The model was given to students because it was thought unlikely the students could develop the model independently, and the model was presented without derivation or formal proof because the mathematics involved is too complex for high school students. As an alternative students proved the sample model's utility and legitimacy using a frequentist approach with a Fathom simulation. Such an emphasis on utility and application lay somewhere between an informal approach that was designed to cultivate intuitions and the more formal mathematical approach appropriate at senior school. The real-world scenario chosen was a public opinion survey of a local controversial and well-publicised issue of supporting or opposing the construction of the Mt. Wellington Cable Car. This phase largely provided the classroom tuition for Research Question 2.

Phase 4. Conduct post-study assessment. The post study assessment was conducted in three parts and under traditional examination conditions. The first part considered students' use of Fathom and their ability to assemble a basic simulation, the second considered students' development of understanding of re-sampling and the sample size model, and the third was a follow-up test item that assessed students' long-term retention of the concepts through re-presentation of the national and state election survey item introduced first in the pre-testing in Phase 1.

The classroom work samples in each of the four phases are summarised in Tables 3.3 – 3.6, the work samples are presented separately in a consistent format that identifies the items' key objectives provides a description of the item, and gives an explanation of the methodology used to analyse students' responses. The work samples are grouped by the three research questions in Tables 3.8 – 3.10.

3.3.2 Phase 1: Pre-study items and introductory exploratory data activity

The researcher was unfamiliar with the students so it was essential to assess students' background skills thought essential for the study. If the study group was unable to demonstrate competency in fundamental skills, the objectives of the study would have to be modified accordingly. The work samples for Phase 1 are presented in Table 3.3.

Table 3.3.
Phase 1: Pre-study Assessment Items and Work samples

Item or Task	Section	Principal objective	Res Q.	Results Chap. 4	Appendix
Basic mathematical skills (Pre-test Q. 1 a-h, Q. 5)	3.3.2.1	Basic mathematical skills thought essential for the study		4.2.1, 4.4.2, & 4.5.3.1	A.1 (Pre-test Q. 1 a-h, Q. 5)
Physical die (Pre-test Q. 4)	3.3.2.2	Students' interpretation of the data of 30 rolls of a physical die	Q. 1	4.2.2	A.1 (Pre-test Q. 4)
Data spread of a class set of a multiple coin toss. (Pre-test Q. 2 & 3)	3.3.2.3	Students' naive understanding of distribution of proportion of heads from 35 trials of a 50 tosses of a coin	Q. 1 & 2	4.2.4	A.1 (Pre-test Q.2 & 3)
Sample size for a national and state election survey (Pre-test Q. 6 & 7)	3.3.2.4	Students' naïve understanding of sample size for large populations	Q. 2	4.2.5	A.1 (Pre-test Q. 6 & 7)
NY Marathon – introduction to Fathom	3.3.2.5	Introduce and use Fathom, the interpretation of a single distribution	Q. 3	4.2.6 & 4.5.4.1	A.2

3.3.2.1 Basic mathematical skills test

The basic skills test sought to establish a baseline of students' number-sense of fractions and percentages, their ability to use functions of the form of the sample size model, and to assess their ability to analyse and interpret dot plot graphs.

The data for Q. 1 were analysed by the proportion of male and female students that gave a correct response, an incorrect response, or no response for each the item Q. 1 (a – h). The data for the physical die, Q.4, were analysed by categorising students' responses about whether the die was fair or unfair and what criteria were used. The data for item Q. 5, the Female race-times, were analysed using the SOLO model. To provide a relational response students identified correctly the fastest race time, provided an appropriate range of "most" of the data, located the centre of the data distribution, and identified the data centre correctly as median, mean, or mode.

3.3.2.2 Physical die

The physical die item sought to establish students' ability to interpret a histogram of the frequency with which each face occurred in 30 rolls of a die, and to determine whether the student considered the die as fair or unfair. The data for the Physical die, Q.4, were analysed by categorising students' responses about whether the die was fair or unfair and what criteria were used to determine that response.

3.3.2.3 Data spread of a class set of a multiple coin toss

This item sought to establish students' sense of the distribution of data of a class set of the proportion of heads of 50 tosses of the coin; to determine familiarity with the +/- notation used to quantify error and accuracy; and to determine students' personal definition of the term "most" as an informal measure of a 95% confidence interval.

The context was a familiar classroom scenario where thirty students tossed a coin fifty times, calculated the proportion of heads, and pooled the data as a class set, but the task was varied by students considering the range where "most" of the proportion of heads would occur. Both parts of the item were posed as multiple-choice questions. The item was also a companion task to the 50 & 500 coin toss where students compared their intuitions with the data generated by a Fathom simulation (Section 4.4.7).

The data were analysed by determining the proportion of students who chose each one of the five alternatives offered and by pairing each individual student's choice of range of the proportion of heads with their choice of the number of students in the class they thought that would have that result (e.g., 28 students in a class of 30). Students' choices were then compared with the theoretical distribution that would occur by chance.

3.3.2.4 Sample size for a national and state election survey (Pre-test)

The context of the task was an opinion poll conducted prior to a national and state election that gave a choice for one of two major political parties. The opinion survey was likely to be contextually familiar: the national election that brought a change in government was newsworthy and held approximately nine months before the study, and although the students were not old enough to be eligible to vote, they would be eligible to vote at a subsequent election. The item first asked students to choose the sample size necessary for a national election of a voter population of 15 million, and second to choose a sample size for a state election of a smaller voter population of 1/10th or 1.5 million voters. The item was offered as a multiple choice. Students could choose a sample size of 10% of the population or choose from a series of three absolute sample sizes presented in decreasing order of magnitude. The second part of the two items asked students to choose a sample strategy from the strategies offered, or otherwise state the strategy that they had used.

The data were analysed first by calculating the proportion of students that had chosen each sample size for the national election, or had not given a response. Second, the data were analysed for whether students had used a consistent or inconsistent sample strategy for both the national and state opinion poll. Third, the data were analysed for the sample size strategies students had used.

The item was subsequently re-presented to students as a follow-up test two months after the conclusion of the classroom study to assess students' long-term development of understanding of sample size.

3.3.2.5 New York Marathon – introduction to Fathom

This lesson was students' first opportunity to use Fathom. An exploratory data analysis task was used because this activity was thought more familiar to students than simulation activities used subsequently. A data set of marathon race times was chosen as providing a familiar context. The students worked in pairs in a peer-tutored approach: one student from each pair was taken to a separate room and fabricated a die using modelling clay that was to be used in subsequent die simulation activities; the other student in the pair was instructed by the researcher in the basic features and use of Fathom with the intention that this student would instruct the other student in the pair. The peer-tutored approach was designed to focus the "tutor" students' attention during instruction, provide a range of learning approaches, create an opportunity for students to apply what they had learnt, and promote collaborative learning.

The researcher introduced the most basic terms and working tools necessary to use the software the Fathom. The key Fathom terms of case, attribute, and collection along with the more familiar equivalents were introduced and defined, and the use of software modules of collection, table, and graph were demonstrated. Ten minutes were allocated for instruction, a time chosen as a reasonable endurance for students of this age-group and which allowed sufficient time for the other student in the pair to complete the practical task of fabricating a die. On returning for peer-instruction, the student was instructed to "explain to the other student in the pair what was demonstrated to you." A guided worksheet was purposefully not provided principally to focus students' attention and to commit basic skills to memory. After approximately five minutes the researcher then drew the class's attention to the assessment task.

Students were instructed to choose a representation of the data, such as a graph, and to write a brief description of their chosen representation. Webbing was provided by a whole class discussion, in which 6-8 questions were developed to guide examination of the data. The researcher emphasised that with Fathom "It is easy to create a graph, but it much harder to create a graph that tells a story." Students' responses provided a base-line for the verbal description of graphical representations of data and the GICS framework (Section 3.2.8) to be introduced subsequently. The lesson plan and the lesson MS-PowerPoint presentation are attached as Appendices A.20 and A.31.

The items were analysed using the SOLO model, and exemplars of students' work at each of the five SOLO levels are provided (Section 4.2.5). In a subsequent section (4.5.4.1) students' development of use of Fathom is examined using the instrumental genesis framework.

3.3.3 Phase 2: Develop acceptance of the Fathom simulation

Phase 2 of the study modelled a statistical enquiry of the fairness of three dice, cultivated an environment of statistical process where decisions were evidence-based, developed a formal notion of a statistic as one number representative of a more complex situation and created a statistic appropriate to the context, and introduced virtual simulation of a familiar random process of tossing a die.

This phase of the study was a series of three iterations that corresponded to formal testing of the fairness of three dice identified as a home-made die, a factory-made die, and the Fathom die. The home-made die identified the die fabricated by students using modelling clay, the factory-made die was a conventional commercially available die, and the Fathom die was the virtual Fathom die simulation. Students rolled the die thirty times and recorded the frequency with which each face occurred. The sample size of 30 was chosen by the researcher because the sample size was divisible neatly by six and consequently gave an integer value (five) for the expected value. The sample size had no mathematical justification, but the sample size was practicable in class, was not excessively intrusive on class time, allowed three dice to be tested, helped sustain students' interest with little risk of classroom management issues occurring. The choice of sample size was used as a topic for class discussion. Comparison of three data sets was designed to shift the focus from mere analysis of data to a task in which students must form an opinion based on the data available (Table 3.4).

Table 3.4.
Phase 2: Develop Acceptance of the Fathom Die Simulator Work Samples

Worksheet & task	Section	Principal objective	Res. Q	Results Chap. 4	Appendix
Home-made die	3.3.3.1	Create and test a home-made die	Q. 1	4.3.2	A.3, Q. 3
Develop a fairness measure	3.3.3.2	Develop a formal objective measure of a die's fairness	Q. 1 & 2	4.3.3	A.3, Q. 5
Fairness measure homework	3.3.3.4	Promote deeper mathematical understanding of the fairness measure	Q. 1 & 2	4.3.4	A.5, Q. 1-3
Fathom virtual die – first Fathom simulation	3.3.3.5	Assemble and test a Fathom virtual die	Q. 1 & 3	4.3.5	A.6
Compare three dice using GICS	3.3.3.6	Final assessment compare distributions using GICS	Q. 1 & 3	4.3.6	A.7

3.3.3.1 Home-made die

The activity sought to stimulate student interest by fabricating a home-made die using Sculpey modelling clay, test the fairness of the home-made dice, and use the observed behaviour of the die as a stimulus to consider the fairness of the die informally.

The activity, where students fabricated and tested a home-made die, was a novel variation of dice activities used widely at school. The activity was based on one developed by Key Curriculum Press (2007, p. 271). The die that students fabricated was identified as the home-made die, a title chosen to convey that the standard of manufacture cannot be equal to that of a commercially manufactured die, and to distinguish it from the factory-made and Fathom dice used subsequently. The die was likely to be unfair or biased, and this provided the stimulus to consider the fairness of the home-made die. The worksheet included questions to prompt students' consideration of fairness of the die.

The data from this item were analysed by identifying the criteria students used when they considered the fairness of the die.

3.3.3.2 Develop a fairness measure

The key objective was to extend students' informal notions of a die's fairness to the development of formal measurement of fairness of a die. Students were asked to propose a formal measurement of a die's fairness. It was emphasised to students that just having a feeling was not sufficient: the class needed to develop a single number – a statistic – that measured the fairness of the die objectively.

Students' proposals for a fairness measure provided a topic for the classroom discussion. After classroom discussion the fairness measure used was the sum of the difference between observed and expected frequency:

$$\sum_{face=1}^{6}(\text{observed frequency} - \text{expected frequency}) \text{ (Equation 2).}$$

Students calculated the fairness measure for the home-made die and repeated the test and fairness measure calculation for the factory-made die. The fairness measure data from the two physical dice – each student contributed a datum point for each – were displayed as a class data set on a poster-size dot plot at the front of the classroom. The dot plots acted as a boundary-object (Star, 1989) for whole-class discussion.

The data were analysed using SOLO by categorising students' proposals for a formal measurement of a die's fairness as descriptive, formal mathematical oriented towards a single statistic, or no response. A response that was descriptive was considered unistructural and a response that included a mathematically oriented aspect was considered multistructural.

3.3.3.3 Whole class discussion

Four topics were used for whole-class discussion with the researcher and the colleague teacher. Students were prepared for the topics by questions posed in the worksheet.

The first topic was the sample size used to test the dice; that is, the number of times the die was rolled. The choice was largely one of convenience and had no mathematical justification.

The second topic of discussion explored the informal criteria students used to consider the fairness of dice. Prompting questions on the worksheet included: Do you think your die is fair? What in the data makes you think your die is fair or unfair? Or are you not convinced either way?

The third topic of discussion was the development of a statistic to measure the fairness of the home-made die. The term measure was used in the study in preference to statistic to be consistent with the terminology used in Fathom, and the statistic was given the title of the fairness measure. The fairness measure, used in the study after classroom discussion that considered students' strategies, was calculated as the sum of the difference between the observed an expected.

The fourth topic examined and compared the fairness measure data generated by the two physical and the Fathom virtual dice, The discussion had several iterations as data for each of the three dice were generated, and stimulus for the discussion – a boundary object – was provided by the data displayed as poster-size dot-plots at the front of the classroom (Figure 4.5). This discussion introduced the GICS framework. Students were asked to identify significant features of the fairness measures dot plots, and the GICS framework was used to categorise the features as either a Global, Individual, Measure of Centre or Measure of Spread feature.

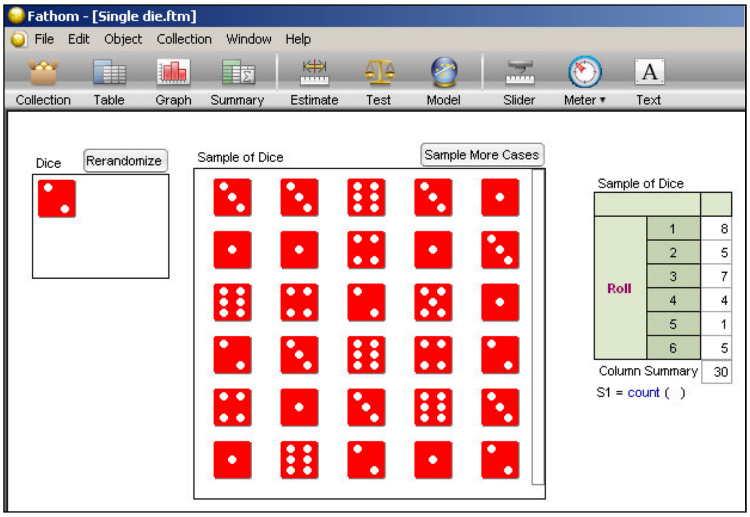
3.3.3.4 Fairness measure homework

The objective of the item was to cultivate a rigorous mathematical approach to the fairness measure and promote deeper understanding by exploring the formal mathematics within the fairness measure. Four tasks were presented: (a) calculate the fairness measure from a graphical representation and identify the least and most fair die, (b) calculate the minimum and maximum values of the fairness measures for a sample size of thirty, (c) reverse the traditional calculation by generating a data set from a given fairness measure, and (d) demonstrate the connection between the graphical representation and the fairness measure. The four tasks were assessed as either correct or incorrect.

3.3.3.5 Fathom virtual die – first Fathom simulation

This activity introduced students to Fathom simulation. Students assembled a Fathom die simulation, checked the simulation as the simulation was assembled using a step-by-step proving process, and tested the fairness of the die simulation using the same procedure used for the physical dice.

This activity was the students' first use of the Fathom die simulator, but their second use of Fathom. Students were presented with a Fathom simulator represented as a die icon displaying a single face of the die. Guided by an illustrated worksheet (For an example see Appendix A.6) the students took a sample, changed the sample size from the default sample size of 10 to a sample size of 30 required for the test, and created a summary that displayed the frequency at which each face occurred. At each stage of assembly the guided worksheet directed students to test that the stage functioned correctly, and that the data representations were internally consistent. Three simple questions to identify most or least frequent occurrence of a face, whether a particular face occurred more than ten times, and whether any particular face did not appear at all, encouraged students to explore informally the behaviour of the dice. The worksheet concluded with the testing procedure to determine the fairness measure and students adding their own data to the poster-size dot plot displayed in the classroom. Figure 3.2 presents the assembled Fathom die simulation as it appeared to the student. The activity was assessed by the ease with which students assembled the simulation correctly.



Sample of Dice

Roll	1	8
	2	5
	3	7
	4	4
	5	1
	6	5
Column Summary		30

S1 = count ( )

Figure 3.2 Workspace of a fully assembled single die simulation.

3.3.3.6 Compare three dice using GICS

The key objectives of this activity were to examine students' ability to compare multiple data distributions of the fairness measures of the home-made, factory-made and Fathom dice; assess the effectiveness of the Global, Individual, Measure of Centre, and Measures of Spread (GICS) framework (Bill, 2007) as a data analysis tool; and determine students' beliefs of the fairness of the Fathom die.

The assessment task was a formal written assessment of the classroom data collection and whole-class discussion of the fairness measures of the three dice. Students were supported in the class in the lessons before the assessment through informal whole-class discussion and small group teacher-led discussion of the features of the three distributions. Students could use their own notes recorded during the class discussion and the three poster size dot-plots of the fairness measures were displayed prominently at the front of the classroom (e.g., Figure 4.5). A comparison of the three dice was not mathematically legitimate, because the home-made dice were not identical. Students worked independently under traditional examination conditions.

Comparison of three data sets shifted the focus from mere analysis of data to an authentic task in which students had to form an opinion and make a decision based on the available evidence. The item was purposefully posed in two parts to encourage students to analyse the data before drawing a conclusion. Students were encouraged orally to use the GICS framework as a check-list. The two classes had generated their own data, so the data, the analysis, and the conclusions differed slightly.

The first part of the task, where students compared the three data distributions, was assessed using the SOLO model. Student exemplars for each of the SOLO levels are attached as Appendix F.1. The second part of the task was analysed in two stages. The first stage categorised students' beliefs of the Fathom die as less fair, fair as, or fairer than the factory-made die, or no belief was expressed and the second stage noted whether or not students argued principally from the available evidence.

3.3.4 Phase 3: Large population sample size model $\boldsymbol{e = \pm 1/\sqrt{n}}$

Phase 3 of the study introduced the large population sample size model and the Fathom coin simulation, examined informally the effect of sample size on the centre and spread of a distribution of measures, modelled statistical enquiry using a frequentist approach and a Fathom simulation to justify the sample size model, and applied the sample size model to contextual tasks of sampling for large population opinion surveys (Table 3.5).

This phase of the study extended students' notions of re-sampling introduced with the fairness measure to the consideration of sample size and the large population sample size model. Students brought to the study informal, and often incorrect, beliefs of sample size, and this study sought to cultivate the development of sound beliefs of sample size. The sample size model $e = \pm 1/\sqrt{n}$, which relates the sample size, n, to the margin of error, e, was introduced to students as a formal mathematical approach to determine sample size.

Table 3.5.

Phase 3 Large Population Sample Size Model e = ±1/√n Classroom Work Samples

Worksheet	Method. Section	Principal objective	Res. Q.	Results Chap. 4	Appendix
The effect of sample size on the fairness measure	3.3.4.1	Transitional activity between fairness measure and sample size activities	Q. 2 & 3	4.4.3	A.8 & A.9
Coin measures 50 & 500 tosses of a coin homework – Part 1	3.3.4.2	Demonstrate understanding of the measures and the two expressions "sample size used to calculate measures" and the "number of measures collected".	Q. 3	4.5.2.2	A.10, Q. 1
Coin measures 50 & 500 tosses of a coin homework – Part 2	3.3.4.3	Construct a measures dot plot using a small sample dot plot as a template	Q. 2	4.4.5	A.10, Q. 2
Physical coin toss (cumulative proportion of heads)	3.3.4.4	Introduce simulations using physical coin	Q. 2	4.4.4	A.11
Fathom virtual 50 & 500 coin toss	3.3.4.5	Assemble a Fathom coin toss simulation. Collect coin measures manually using Fathom	Q. 2	4.4.6	A.12
Compare intuition of a 50 coin toss with a Fathom coin toss	3.5.4.6	Compare naïve sense of distribution of proportion of heads of a 50 coin toss with a Fathom simulation	Q. 2	4.4.7	A.12
Mt. Wellington cable-car (naïve)	3.3.4.7	Contextual task using sample size model (Homework)	Q. 2	4.4.8	A.13
Large population sample size model	3.3.4.8	Introduce and provide evidence for sample size model. Use Fathom to generate the data.	Q. 2	4.4.9	A.14, Q. 2
Opinion polls	3.3.4.9	Whole-class discussion	Q. 2	None	None

3.3.4.1 The effect of sample size on the fairness measure

This activity provided the opportunity for students to examine the effect of sample size on the fairness measure, modify the fairness measure to a percentage measure calculated as the difference between the percentage expected (16.7%) and the percentage observed, and use a Fathom die simulation to collect the percentage fairness measures at sample sizes of 30, 300, and 3000.

This was the first activity that examined specifically the effect of sample size – a key theme of the research study. Students re-assembled the Fathom die and used initially the same sample size of 30 as in the previous physical and Fathom simulation activities. Increasing the sample size to 300 showed that a fairness measure based on frequency would not allow comparisons to be made at different sample sizes – the measure must be re-expressed as a percentage of rolls where any face appeared. Students used the Fathom die simulation to collect the raw data (roll the dice), but calculated the percentage fairness measures at the three sample sizes of 30, 300, and 3000 manually. The individual fairness measures were entered into the researchers' computer that displayed the class set of data as a dot plot projected onto a whiteboard at the front of the class. This was designed to mimic the manually generated measures dot plots in earlier activities.

To complete the activity correctly the students copied the class generated data to their own worksheets and completed a series of questions designed to highlight features of the three graphs. Students were asked to locate the centre of the data, describe verbally the centres and spreads of the three distributions, and use the graphs to interpolate the fairness measure to two other sample sizes. Two additional questions asked students to consider the relationship between sample size, the percentage fairness measure, the fairness of the die, and the natural variation. This activity concluded the use of the Fathom die simulation in class.

The task was a collaborative, highly supported, classroom activity with a detailed worksheet, teacher-led discussion, and no formal assessment task, but students' completed worksheets were used for analysis. The task was challenging, with a substantial increase in complexity and abstraction relative to the previous activities. Modifying the fairness measure from a frequency to a proportion-based percentage fairness measure required a shift from additive to proportional thinking – the previous fairness measure only allowed comparison at a constant sample size. The focus for students' analysis and discussion now lay with the percentage fairness measure, and this measure was not directly connected to the underlying data of the frequency count for each face. The introduction of a percentage or proportional-based measure was also designed to provide a foundation for the coin simulation and the measure of "proportion of heads" used subsequently.

The data from the girls' class were not included in the analysis of the mathematical concepts, because a simplified Fathom worksheet was used with the girls that proved ineffective and the mathematical purpose of the lesson was lost The Fathom worksheet was subsequently modified to the original format that included screen-shots and detailed instructions (Appendix A.8), and this worksheet was used in the boys' class.

The task was analysed using the SOLO model, and exemplars of all levels of boys' responses are provided. Boys' responses to the item Q.6, which explored students' understanding of the mathematical relationship between a die's fairness and sample size, and the natural variation that occurs in chance behaviour, are also presented. The item was subsequently re-examined using instrumental genesis to consider students' development of use of Fathom.

3.3.4.2 Coin measures 50 & 500 tosses of a coin homework – Part 1

The objective of this item was to assess students' understanding of key terminology used in re-sampling of "the sample size used to calculate a measure," and "the number of measures collected." The activity was presented to students as a two-part homework assignment, and students worked collaborative if they wished. Completion of the activity was voluntary, and only the motivated students completed the task. As a research instrument the activity provided evidence of a range of student responses.

Students were presented with a dot plot of the proportion of heads of a 50 coin toss. To complete the task successfully students identified correctly the sample size used to calculate the proportion of heads, provided a meaningful title for the measure of a proportion of heads, identified the number of measures collected by counting the number of measures presented in the dot plot, and gave the expected proportion of heads for a fair coin. The task was analysed by categorising students' responses as either correct, incorrect, or no response for the four elements in the task.

3.3.4.3 Coin measures 50 & 500 tosses of a coin homework – Part 2

The objective of this item was to assess students' ability to construct a distribution of the proportion of heads of a series of 500 tosses of a coin based on the distribution of proportion of heads from a series of 50 tosses of a coin.

The task extended the earlier work on the effect of sample size on the fairness measure of dice simulation to the proportion of heads from a coin simulation. Students were provided with a dot plot containing 30 measures of proportions of heads calculated from a series of 50 tosses of a coin, and this was used as a template for a 500 tosses of a coin. To complete the task successfully students sketched a dot plot that included three features: the data centred on a proportion of 0.5, and a similar number of proportions of head and a narrower spread of the distribution than the 50 tosses of a coin measures dot plot. The item was analysed using the SOLO model using the three criteria that the response for the 500 tosses of a coin included that the centre of the data distribution was located correctly at a proportion of heads of 0.5, a similar number of measures was used, and the spread of the distribution was narrower than for the 50 tosses of a coin.

3.3.4.4 Physical coin toss (cumulative proportion of heads)

The objective was to use a physical coin simulation to provide a foundation for the subsequent introduction of the Fathom virtual coin, calculate the proportion of heads, and calculate the difference between the observed and expected proportion of heads. Students worked in pairs; they tossed a fair coin 50 times, recorded the sequence of heads and tails that occurred, made a running tally of the frequency of heads, calculated the difference between observed and expected, and recorded the cumulative proportion of heads on a trend graph. The task examined students' informal observations of the relationship between the differences between observed and expected frequencies, and sample size.

3.3.4.5 Fathom virtual 50 & 500 tosses of a coin simulation

Students assembled a Fathom coin simulation that included a summary and table using the same procedure used for the Fathom die, tested the simulation using a progressive step-wise process, used the formula editor to compute the Proportion of Heads, and completed the transition from physical to virtual simulations.

To support students' transition from physical to virtual simulations the first Fathom coin simulation was used in the same class period as the physical coin activity. The Fathom simulation was assembled in an identical process to the earlier die simulations using detailed worksheets with screen-grabs, detailed instructions, and a step-wise process that ensured the simulation operated as intended. Students' understanding of the key terminology of “sample size used to calculate a measure” and “the number of measures collected” was checked. Students used the Fathom simulation to generate 30 measures of “proportions of heads” at two sample sizes of 50 and 500 tosses of the coin, but the data were collected and recorded manually on dot plots. Students were asked a series of five questions identifying the key features of the graphs including the expected value and the effect of sample size on the measure of proportion of Heads.

Students’ responses were assessed using the SOLO model. To demonstrate a relational response students assembled the simulation, collected measures of the proportion of heads at the two sample sizes of 50 and 500, recorded the data on two dot plots, identified a meaningful centre of the data, and noted that increasing the sample size both reduced the spread of the data and tended to shift the centre of the data to the expected value of 0.5 proportion of heads.

3.3.4.6 Compare intuition of a 50 tosses of a coin with a Fathom coin toss

The item provided students with an opportunity to compare their intuitive understanding of the distribution of the proportion of heads of a 50 tosses of a coin established in the pre-test Q. 3 with a Fathom simulation. This item was an extension of the previous task and used the data of the proportion of heads collected for a sample size of 50 tosses of a coin. Students marked the Fathom 50 tosses of a coin measures dot plot to indicate the range they defined as “most” in the pre-test, counted the number of proportion of heads measures that occurred within that range, and compared the data with their own intuitive response that they gave in the pre-test.

Students’ responses were assessed in two parts. For the first part students were assessed as over-estimating the number of measures that occurred within a range, making a prediction that was similar to what was observed, under-estimating the number of measures that occurred, or giving no response. The second part considered if the student had assessed the evidence objectively, or what other criteria were used.

3.3.4.7 Mt. Wellington cable-car (naïve)

The principal objective of the item was to examine students’ consideration of the sample size used for a large population public opinion survey of a familiar local contextual issue, and to provide a base-line to assess students’ developmental understanding of sample size of contextual tasks.

The Mt. Wellington cable-car task was a public opinion survey of whether the Hobart community (population 200,000) was in favour, or against, a proposal to install a cable-car on Mt. Wellington, Hobart. The proposal to construct the cable-car was a long-running and controversial local development issue that had received significant media attention, so it was likely that students were aware of the issue. The task was designed to be experiential and placed the student as a professional city council planning officer responding to criticism that the survey conducted was inaccurate because the sample size of 900 used was too small. The task was presented to students as a homework item, included as part of whole-class discussion, and presented again to students in the post-study test as a formal assessment item.

The task was assessed using the SOLO model. To complete the task successfully and provide a relational response students calculated the margin of error using the large population sample model $e = \pm 1/\surd n$, related sample size to accuracy, noted that a sample size was independent of population size when sampling from large populations, noted the importance of a random and representative sample and that a sample is an imperfect representation of a population, rejected 10% of population as an unfeasible sample size, and considered the practicalities and cost of conducting survey.

At this point of the study students were unlikely to have exposure to many of these concepts, so low level responses were anticipated. The same assessment criteria were used in both the initial homework item and the final assessment task to provide an assessment of students' development.

3.3.4.8 Large population sample size model $e = \pm 1/\sqrt{n}$

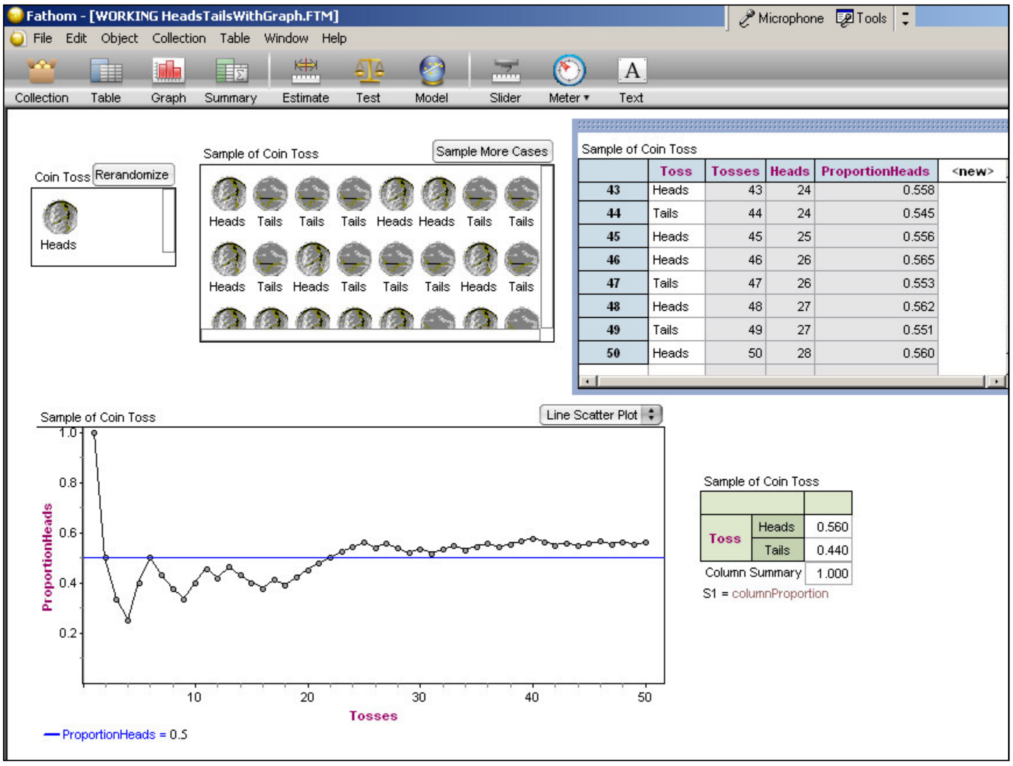
This task introduced the large population sample size model $e = \pm 1/\sqrt{n}$ as an estimate of, e, the margin of error for a sample size n, in chance processes; calculated manually the margin of error for sample sizes of 50, 100, 400, 900, and 1600; used a process of statistical enquiry to justify the sample size model by comparing the model with the data generated by a Fathom simulation; and introduced the Fathom formula editor and Fathom's number formatting feature. This activity extended the previous activity that sought to demonstrate that the spread of measures decreased as the sample size increased to formally quantifying the spread of measures. This activity also provided an introduction to confidence intervals studied at more senior school years.

The activity introduced the sample size $e = \pm 1/\sqrt{n}$ rule as an estimator of the difference between the observed and expected values. More formally the model calculates a 95% confidence interval, or in the terminology of the study the margin of error. The model predicts that approximately 95% of all simulation results will be bounded by the rule, so in 30 simulations used typically in class approximately one or two results would be expected to occur outside of the margin of error.

Students calculated the margin of error manually at a sample size of 49, 100, 400, 900, and 1600 – sample sizes chosen because the margin of error is calculated conveniently. A formal derivation of the model was inappropriate for Year 9, so the model's accuracy and utility were confirmed using a frequentist approach and a Fathom simulation. Students tested the model by running the Fathom simulation six times for each of the five sample sizes in sequence and then compared the calculated results using the model with data generated by the Fathom simulation. Students first assessed the data their own simulation had generated, and then in whole-class discussion considered the data generated by the other students in the class.

In contrast with previous activities the Fathom simulation was presented to students fully functioning (Figure 3.3), but students modified the simulation to include two additional features: the formula editor was used to calculate the difference between the observed value and the expected value of 0.5, and the format feature was used to present the data to two significant places. The fully-functioning simulation was designed to allow as large a proportion of the class time as possible to examine the mathematical concepts.

Students' responses were assessed using the SOLO model. To provide relational responses students calculated the margin of error for the five sample sizes, assembled and used the simulation, and, on the basis of the available evidence, determined whether the sample size model was a reasonable estimate of the margin of error.



Sample of Coin Toss

	Toss	Tosses	Heads	ProportionHeads	<new>
43	Heads	43	24	0.558	
44	Tails	44	24	0.545	
45	Heads	45	25	0.556	
46	Heads	46	26	0.565	
47	Tails	47	26	0.553	
48	Heads	48	27	0.562	
49	Tails	49	27	0.551	
50	Heads	50	28	0.560	

Sample of Coin Toss

Toss	Heads	0.560
	Tails	0.440
Column Summary		1.000

S1 = columnProportion

Figure 3.3. Fathom simulation as presented to students.

3.3.4.9 Public opinion surveys (whole-class activities)

The objective of these activities was to examine the sample size for a public opinion survey. The activities included whole-class discussion and researcher demonstration of a Fathom simulation of a public opinion surveys with response of either For or Against. The whole-class discussion used, as a boundary object, a Fathom simulation to demonstrate margin of error and the likelihood of a survey producing a result counter to the underlying population. The activities were not assessed and no data were collected.

3.3.5 Phase 4: Post-study assessment

This assessment comprised three parts: a formal written examination paper, a test where students assembled a basic Fathom coin simulation, and a follow-up test where the National and state election item was re-presented. Students worked independently under traditional examination conditions. The work samples are presented in Table 3.6.

Table 3.6

Phase 4: Post-study Assessment Items and Tasks

Worksheet or item	Method. Section	Principal objective	Res. Q.	Results Chap. 4	Appendix
50 students in a Year 9 maths class	3.3.5.1	Assess understanding of key terminology of "sample size used to calculate a measure" and "the number of measures collected"	Q. 3	4.5.2	A.15, Q. 1
Federal election survey: Howard and Rudd	3.3.5.2	Interpret the accuracy of a public opinion survey for a given sample size	Q. 2	4.4.10	A.15, Q. 2
Mixed up measures dot plots	3.3.5.3	Place measures dot plots of three sample size in correct sequence	Q. 3	4.5.3	A.15, Q. 3a
Mathematics of the sample size model	3.3.5.4	Complex tasks with sample size function	Q. 2	4.4.11	A.15, Q. 3b & c
Badly biased coin	3.3.5.5	Measures dot plot of a biased coin	Q. 3	4.5.3	A.15, Q. 4
Mt. Wellington cable-car	3.3.5.6	Contextual task applying sample size model. Companion task to 3.5.4.6	Q. 2	4.4.12	A.15, Q. 5
Fathom basic skills test	3.3.5.7	Test Fathom procedural skills	Q. 3	4.5.4	A.16
National and state election (Follow-up test)	3.3.5.8	Assess students' long-term development of understanding of sample size. Companion task to 3.3.2.3	Q. 2	4.4.16	A.17

3.3.5.1. 50 students in a Year 9 maths class (Post-study assessment Q. 1)

The objective of this item was to assess students' understanding of the measure of the proportion of heads of a coin toss by asking students to provide a meaningful title for the measure, and to identify correctly the sample size used to calculate a measure and the number of measures collected. This item was assessed as the proportion of male and female students who provided a meaningful title for the measure, identified correctly or incorrectly the sample size used and the number of measures collected, or provided no response. Incorrect responses were then further classified to identify common themes or errors.

3.3.5.2 Federal election survey: Howard and Rudd (Post-study assessment Q. 2)

This item assessed students' understanding of the relationship between survey accuracy and the sample size used. The 2007 Australian Federal election led to the defeat of incumbent Prime Minister John Howard by the Opposition Leader Kevin Rudd. The item was designed to assess students' ability to interpret the result of an opinion survey that they might encounter in the media. The item was purposefully presented in the final assessment prior to the mention, and the cue, of the sample size formulae.

The item was assessed using the SOLO model. To complete the task successfully students either demonstrated or recalled the margin of error for the sample size of 1600, related the accuracy to the result, and stated the outcome "as likely but not certain."

3.3.5.3 Mixed up measures dot plots (Post-study assessment Q. 3 (a))

The objective of this item was to evaluate students' understanding of graphical representations of re-sampling data described in this study as "measures dot plots." To complete the task successfully students placed the three dot plots in the correct ascending order of sample size. Students were assessed as placing the three measures dot plots in the correct sequence, reversing the correct sequence, confounding sample size if no clear sequence was given, responding incorrecy unclassified if the intent was unclear, or giving no response.

3.3.5.4 Mathematics of the sample size model (Post-study assessment Q. 3 (b & c))

This item sought to assess students' ability to manipulate the sample size function $e = \pm 1/\sqrt{n}$ and to demonstrate an understanding of the formal mathematics of sample size. The first item, (3b), varied the language of the question slightly and asked students to calculate the range rather than margin of error within which most results would occur for a sample size not used previously of 200,.. The correct answer was 0.07. The students were assessed as providing workings that were fully correct, partially correct, incorrect, or providing no response. The second item, (3c), reversed the conventional order of the calculation and asked students to calculate the sample size for a range of a collection of measures that were presented on dot plot. This was annotated as a "tough" question so that students were not discouraged. The correct response was a sample size of 100. The students were assessed as providing workings that were correct, incorrect, or not giving a response.

3.3.5.5 Badly biased coin (Post-study assessment Q. 4)

The item sought to evaluate students' ability to construct a measures dot plot for a coin biased towards heads. To complete the task successfully the dot plot should have a scale and the data distribution centred between a proportion of heads between 0.5 and 1.0. Students' responses were assessed using the SOLO model..

3.3.5.6 Mt. Wellington cable car (Post-study assessment Q. 5)

This item sought to assess students' development of understanding of sample size when sampling from a large population, and assess students' ability to apply the large population sample size model $e = \pm 1/\sqrt{n}$ and interpret the result. This was the same task presented to students as the homework item (Section 3.3.4.7) to determine their naïve response. Re-presenting the item was designed to assess students' development of understanding of sample size as a consequence of the study.

The task was assessed using the SOLO model. To complete the task successfully students calculated the margin of error using the sample model, related sample size to accuracy, noted that a sample size was independent of population size when sampling from large populations, noted the importance of a random and representative sample and that a sample as a imperfect representation of a population, rejected the 10% of population sample size model as unfeasible, and considered the practicalities of conducting a survey. The task used a second cycle SOLO analysis because to complete the task students needed to firstly understand the sample size model, and secondly apply the model to a contextual task (Campbell et al., 1992).

3.3.5.7 Fathom basic skills test

The objective was to assess students' ability to assemble and use a basic Fathom coin simulation independently. The study used Fathom software principally to promote learning of mathematical concepts, and the development of technical skills in using Fathom was not a priority. Nevertheless, students' development of basic skills in the use of Fathom may allow students to attend to the mathematical concepts under study without substantial distraction. Students were assessed as assembling the simulation correctly, partially correctly, or giving no response.

3.3.5.8 Sample size for National and state election survey (follow-up test)

The test was designed to determine whether sustained long-term development of understanding of sample size had occurred, because it was these beliefs and understandings that students will take into the wider community outside of school. This item was the same item used in the pre-test (Section 3.3.2.4), and it was re-presented to the students two months after the conclusion of the study. The two colleague teachers administered the test and mailed the completed tests to the researcher.

The data were examined and analysed in the three stages. First, the proportion of students that had chosen each sample size for the national election, or had not given a response was calculated. Second, whether students had used consistent or inconsistent sample strategies for both the national and state opinion poll was determined. Third, what sample size strategies students had used were identified.

3.4 Students' post-study questionnaire and test

The students' post-study questionnaire considered four themes: the value of Fathom as a software learning tool; students' understanding of concepts considered fundamental to the use of simulation such as key terminology; students' understanding of concepts of sample size when sampling from a large population; and students' attitudes to the study. The thirty questions presented to students were coded on a five-point Likert scale, and their responses were subsequently consolidated to the three responses of strongly agree / agree, neutral, and disagree / strongly agree because a three-point scale was thought to provide sufficient detail. The questionnaire is attached as Appendix A.18.

Confidence in the randomness of the Fathom virtual die simulation was examined in Items 7, 12, 14, 21, and 25 through such items as 7: "By the end of the unit I was convinced the Fathom coin was as random as it needed to be for the work we were doing." Students' understanding of sample size was examined in Items 10, 11,13, 17, 19, 20, 23, 26, 28, and 29 through items such as 13: "Doubling the sample size halves the error." Students' use of Fathom and Fathom simulation including key terminology were examined in Items 1, 3, 4, 5, 6, 9, 14, 15, 16, 18, and 29 through such items as 6: "I was confused by 'sample size' and 'the number of measures calculated'." General attitude to the teaching sequence and pedagogy were explored through Items 2, 22, 24, 27, and 30 through such items as 22: "I would rather solve maths problems which had one clear answer." Students' responses were analysed as the proportion of students that provided each category of response. The responses of the boys and girls were then compared.

3.5 Researcher's professional journal

The professional journal provided the opportunity to describe the research study, the social interactions, and the physical limitations of the learning environment. It also provided an opportunity for self-reflection of professional practice, as well as providing a source of data for the study, and to record common behaviours and emerging themes. An essential component of the research study was the cultivation of classroom discussion and the development of socio-mathematical norms within the classroom.

An audio recorder was worn by the researcher in class, but this was able only to record one-on-one conversations that occurred in the class. The colleague teachers' casual observations during the lessons and the lesson review were included in the journals. The journal was a reference for whole-class discussion as it was impracticable, and probably intrusive and counter-productive, to record fully all the class discussions in an electronic form. The journals provided anecdotal evidence only. Transcribed copies of the professional journals are attached in Appendix E.

3.6 Colleague teachers' post-study interviews

The concluding interviews conducted with the two colleague teacher explored the themes of the teacher's professional backgrounds and their perception of the student cohort, the appropriateness of the topic of study for Year 9 students, students' acceptance of the simulation as legitimate, whole class discussion to promote understanding, the research pedagogy, and the value of Fathom as education software..

Kvale (1996) describes one method of analysis of an interview as a three step procedure of transcription, clarification, and analysis proper. Clarification is a process of eliminating superfluous material, such as digressions and repetitions, and identifying material that is pertinent for the study. The interview analysis proper used the approach of meaning condensation, where the interviewee's responses as understood by the researcher are re-expressed as simply as possible (Kvale, 2007). The themes and the condensed meanings of the two colleague teachers were compared, and extracts of the interview were quoted to support this thesis.

The interviews were conducted separately and in private. The interview with the colleague teacher of the all-girl class was conducted after the conclusion of the study, but the interview with the colleague teacher of the all-boy class was postponed six months because the teacher took extended leave.

The interview protocol, the analysed transcripts, and a summary of the transcribed interviews categorised by themes are attached as Appendix C1 – C4.

3.7 Detailed study research items

3.7.1 Overview

The principal objective of the detailed study was to re-examine and explore issues identified in the classroom study that students found challenging or otherwise significant and to provide information to complement the classroom study. Six pairs of students, all of whom had participated in the classroom study, participated in the detailed study. The detailed study was conducted as a 20 minute post-study interview conducted at the conclusion of the classroom study and a 90 minute workshop conducted at the university office of the researcher about six weeks after the classroom study. The tone adopted was easy familiarity; formal assessment was completed, and students were encouraged to speak freely to identify "what worked, what didn't."

Five aspects were identified for examination in the detailed study: the first four aspects were analysed using situated abstraction, and the fifth, students' development of use of the Fathom software, was analysed using the instrumental genesis.

- The acceptance of the Fathom virtual die, and the preference of the virtual die relative to a conventional physical die.
- The acquisition and use of terminology for re-sampling of including sample size, "sample size used to calculate a measure" and "the number of measures collected."
- The interpretation of numeric representations of data, and how students may be supported in making the transition from additive to proportional and distributional data representations.

- The interpretation of the measures dot plots, and a comparison with the more familiar cumulative proportion trend plot.
- The development of the procedural use of Fathom. The tasks considered students' recall of basic operations and the degree of support required to complete basic tasks.

To support examination of the five aspects students were presented with a series of eight activities grouped into two sets of four tasks; the first four tasks examined students' recall of the use of the basic features of Fathom, and the second set of four tasks examined re-sampling and explicitly quantifying sample size. The eight tasks are presented in Table 3.7.

Table 3.7.
Detailed Study Work Samples

Task	Method. Section	Principal objective	Res. Q.	Results Chap. 4	Appendix
Part A: Die one face only, or six faces presented simultaneously	3.7.2	Orientate students to the workshop environment, determine level of support to take a sample, identify preference for die simulation representation	Q. 3	4.5.4.5	D.1, Part A
Part B: Coin side only or both sides presented simultaneously	3.7.3	Identify criteria students use to determine an extreme event in ten tosses of a coin, and the level of support required to create a Summary	Q. 3	4.5.4.5	D.1, Part B
Part C: The effect of sample size on preference for data representation	3.7.4	Determine students' preference for a data representation and whether the preference is affected by sample size.	Q. 2 & 3	4.3.3 & 4.5.4.2	D.1, Part C
Part D: Three potentially biased virtual dice	3.7.5	Identify the sampling strategies and sample sizes used to determine whether a coin is biased	Q. 1 & 3	4.5.4.2	D.1, Part D

Table 3.7. (cont.)

Detailed Study Work Samples (cont.)

Task	Method. Section	Principal objective	Res. Q.	Results Chap. 4	Appendix
Part E: Graphs and the Law of Large Numbers	3.7.6	Assess students' ability to construct a meaningful graph independently, determine whether a fixed reference line supports interpretation, level of support required to assemble a graph	Q. 3	4.3.5	D.1, Part E
Part F: Cumulative proportion of heads graph and sample size	3.7.7	Extend use of cumulative proportions of heads graph to the interpretation of sample size and accuracy	Q. 2 & 3	4.3.6 & 4.5.3.7	D.1, Part F
Part G: Measures dot plots at sample sizes of 50 & 500	3.7.8	Examine students' use of "sample size used to calculate a measure" and "number of measures collected." Examine ability to interpret a measures dot plot.	Q. 3	4.5.3.7	D.1, Part G
Part H: For and Against – contextual sampling task	3.7.9	Assess students' ability to apply developing intuitions of sample size to a contextual task	Q. 2	4.5.4.2	D.1, Part H

The first four activities (Part A to D, described in Sections 3.7.2 – 3.7.5) were designed to orientate and place students at ease in the unfamiliar working environment of an office at the university. Fathom was re-introduced to students through a series of basic tasks that examined their personal preferences for numeric data representations and their intuitions of probability. The tasks increased progressively in complexity. The simulations were presented partially assembled; each task re-introduced a feature of Fathom and students were asked to complete the assembly of the simulation. The students had not used the software for approximately six weeks. The activities assessed students' recall of terminology and the procedural use of the software by asking students in Part A to take a sample, in Part B to create a summary (a data table), and in Part C to change the sample size. Students' preferences for specific numeric data representations were established by offering different representations simultaneously – in this instance a coin toss series was presented as the number of heads, the proportion of heads, and the percentage of heads that occurred – and examining how the preferences change with sample size.

The second series of four tasks (Part E to H, described in Sections 3.7.6 – 3.7.9) focused on sampling, sample size, and graphical representations of data generated by a virtual coin toss. Parts E, F and G tested students' recall of constructing a graph including choice of attribute used, and examined their interpretation of the information within the graph. Part F and part G were companion tasks. The two tasks examined multiple series of coin tosses simultaneously, but presented the information in two different graphical formats: Part F presented the data in the traditional graph of the "Cumulative Proportion of Heads" plotted against the frequency of coin tosses, and Part G utilised Fathom and presented measure dot plots of "Proportion of heads" at two different sample sizes. Part H applied the concepts studied to the contextual task of sampling from a large population.

3.7.2 Part A: Die one face only, or six faces presented simultaneously

The simulation was presented as two sub-tasks with one die shown as a single icon only (which arguably most naturally mimics a physical die), and the second die with all six faces shown simultaneously (which may foster confidence that all six faces are indeed present). It was emphasised to the students that the two icons were mathematically equivalent and the choice was entirely one of personal preference; students were asked to identify their preference as "what is most natural to them." The significance of this choice became apparent only when students were presented with the simulation and the sample taken from the collection, which in the instance of all six faces presented simultaneously, may have appeared similar to each other.

The level of support provided to students to take a sample from the simulation was assessed on a scale of completed by "self" / "after demonstration" / "demonstration and support."

3.7.3 Part B: Coin, one side only, or both sides presented simultaneously

The coin simulation was used for the more formal mathematical sampling tasks that were to follow. The coin simulation provided to the students had a coin simulation collection and a sample defaulted to a sample size of ten tosses of the coin. In common with the previous task students were asked to identify the simulation they preferred of a coin either represented as one face only or both faces presented simultaneously. Students created a Summary, which required drag-and-dropping the Summary icon from the taskbar to the worksheet, locating, and drag-and-dropping the appropriate attribute to the Summary. The students then repeatedly ran the simulation until a result occurred that was, in each student's opinion, extreme or odd. The level of support provided to students to create a Summary was assessed on a scale of completed by "self" / "after demonstration" / "demonstration and support," and the proportion of students requiring each level of support was calculated.

3.7.4 Part C task: The effect of sample size on a preference for data representation

The objective of the task was to determine if and how students' preference for the type of numeric data displayed changed with sample size. Students were provided with a coin simulation set to the default sample size of 10 and summaries that presented the data of a coin toss displayed in the three ways simultaneously, as a frequency, a proportion of heads, and a percentage of heads.

The study speculated that at small sample sizes divisible conveniently by two or ten students would prefer results expressed as a frequency, i.e., "same number of heads and tails," and that students would shift to a proportional representation at large sample sizes or samples sizes not easily divisible by two. Students' shift in preference from frequency to proportional representation would reflect a shift from additive to proportional thinking.

Students' preference for a frequency or proportional data representation and the sample size at which any change of preference occurred were recorded.

3.7.5 Part D task: Three potentially biased virtual coins

This activity presented students with three virtual coins that were described as possibly biased. A virtual coin eliminated the labour required to collect data and liberated students to focus on the relationship between sample size and the convergence to the expected long run value.

The activity asked students to recall how they might test a physical coin as an introduction to determine whether a virtual coin was biased. Students were asked to predict the sample size they felt likely to be required, and to use the simulation until they were confident – as distinct from certain – that the coin was biased or not. Any sample size could be used, but students were asked to determine whether the coin was biased using the minimum sample size. The simulations were presented to students fully operating, using different versions so that the two students worked independently, and constructed so that the sample size was increased initially by one each operation. The activity was designed to encourage students to shift their focus from a frequency to a proportional representation of data representation and to make a decision with uncertainty, i.e., based on probabilistic rather than deterministic information.

Students were assessed on the strategy used to determine fairness, whether they preferred data displayed as a frequency or as a proportion, whether they changed the sample increments from the default single increments to larger increments of five or ten, and the sample size used. To manage the computer workspace the three virtual dice were presented sequentially.

3.7.6 Part E: Graphs and the Law of Large Numbers

The activity was a virtual equivalent of a traditional Law of Large Numbers activity of a coin toss, and it was offered in response to the observation in the classroom study that students found measures and the collection of measures challenging. Students were provided with a functioning Fathom die simulation of a coin icon, a sample, and a summary, for which students constructed a graph.

Students used the simulation to complete a series of three sub-tasks: (a) demonstrate that the simulation is understood by operating the simulation and describing it verbally; (b) construct a graph by drag-and-dropping the Graph icon from the toolbar and placing an appropriate attribute (Proportion of Heads was preferred) on the dependent axis; and (c) interpreting the graph verbally. Students were asked to construct a graph of their own choosing, but that the graph should "tell a story" – students needed to be able to construct a graph that they could interpret. This enabled students to demonstrate the existing graph schemes and to allow students' evolution to a more sophisticated response to be observed.

The researcher switched off the default graph auto-scaling feature to eliminate the distraction of the graph appearing to shift each time the simulation was run, and included a reference line on the graph at the expected value of 0.5. The level of support provided to students to create a Summary was assessed on a scale of completed by "self" / "after demonstration" / "demonstration and support."

3.7.7 Part F: Cumulative proportion of heads graph and sample size

This activity was an extension activity of the previous activity (Part E), with the additional complexity of several runs of the simulator appearing simultaneously on the one trend graph. Students were provided with a fully functioning simulation. To complete the activity students interpreted the features on the graph and calculated approximately the difference between the observed and expected frequencies. The Fathom workspace appeared similar to that presented in Figure 3.3.

Students were assessed on their ability to identify and explain the features on the graph, to explain the behaviour of the proportion of heads at sample size at 100 and 600 tosses of the coin, to calculate the difference between observed and expected, and to describe how the graph might be used to relate sample size and the accuracy of a survey.

3.7.8 Part G: Measures dot plots at sample size of 50 & 500

This activity was designed to match closely the method used in the classroom study where Fathom was used to generate proportions of heads measures, but which were graphed manually. The activity varied the classroom activity and Fathom was used to both generate and record graphically the measures. Students were provided with a functioning die simulator, a sample, a collection of measures, and a graphical representation of the measure “proportion of heads.” Students ran the simulation several times. Students were assessed on their ability to provide an oral explanation of the simulation and its features, and to provide a meaningful name of the measure, to identify correctly how many measures were being collected, and the sample size used to collect the measure. Students were then asked to compare the measures dot and the cumulative proportion of heads graphs as methods most effective to them of displaying the accuracy of sampling.

3.7.9 Part H: For and Against – contextual sampling task

The activity concluded the detailed study by providing students with an opportunity to apply their developing understanding of sample size and of virtual simulation to the contextual task of surveying public opinion survey. The population to be surveyed was approximately 200,000 people; this population, although finite, cannot be represented conveniently as individual items data in Fathom directly because the simulation was too slow. The alternative was to model this large finite population as an infinite population in the same process used previously to simulate die and coins systems. It was conceivable that students preferred to see the entire population, and to sample from that population, and simulating the population created an additional abstraction and grounds for rejecting the simulation as not legitimate. Students were presented with a simulation of a collection with iconic representations of the individuals, a sample, and a summary. Students were assessed on their ability to provide an oral description of the simulation, estimate the sample size required and its associated accuracy, and run the simulation.

3.8 Data analysis

3.8.1 Data storage

A database was used as the primary storage of student codes, gender, log of worksheets completed and assessment grades. Information was then exported to a spreadsheet. The handwritten researcher's journal and notes taken during the concluding interview with the colleague teacher were transcribed to a word-processor. All electronic worksheets submitted by students' and interview audio recordings were saved to flash drives. Hard copy worksheets are stored securely at the University of Tasmania. Students' worksheets were scanned and stored as electronic files.

3.8.2 Evaluation of multiple choice or correct / incorrect items

Questions to assess students' ability to calculate fractions and percentages were assessed on a traditional correct or incorrect basis. Multiple choice questions were categorised by response, or marked as either correct or incorrect, or within a hierarchy of most preferred to least preferred alternative. The number of students that gave a particular category of response was recorded as a proportion or a percentage of the class.

3.8.3 Evaluation of student's worksheets using SOLO

The principal instrument of the research study was the student worksheets collected at the end of each lesson. The wide range of tasks and frequent assessment was designed to allow comprehensive observation of student learning of the mathematical concepts. Continuous assessment also accommodated student absences or the occurrence of lower motivation on a particular task or day. The researcher repeated the assessment of the students' responses to ensure that the assessment was stable and consistent. Several of the worksheets were complex and these were evaluated as two smaller sub-tasks or stages.

The assessment process consisted of the following sequence:

- developing a draft SOLO assessment rubric for each task;
- performing a preliminary analysis of the data for an item to establish the range of student responses;
- incorporating the range of responses to complete the descriptors for the SOLO based assessment rubric;
- assessing individual student's responses by continual referral to the SOLO assessment rubric;
- recording student assessment;
- setting the first assessment aside overnight; then
- repeating the assessment;
- comparing the first and second assessment;
- resolving any discrepancies between the first and second assessments by more considered assessment; and
- recording the final assessment.

The SOLO model (Section 2.6.4) incorporates a hierarchy of modes and responses. Students of the age of the cohort in this study were anticipated to work principally in the concrete-symbolic mode with some elements in the formal mode (Collis, 1975). Response, which is an individual's ability to respond with increasing sophistication to the task, is described by a five level hierarchy of prestructural, unistructural, multistructural, relational, and extended abstract responses. The principles of Campbell, Watson and Collis's (1992) second cycle analysis were applied to assess students' development of understanding.

The assessment rubric for each task includes descriptors for each SOLO level. To simplify the rubric the complete descriptor for each task, including students' ability to integrate all aspects of the task, provides the higher level relational response. Descriptors for each SOLO level of response were developed in draft form that was subsequently complemented by the students' own responses to give the assessment rubric used. Students who provided an incomplete list of descriptors and who were unable to integrate all aspects of the task were assessed at a level lower than a relational response. To avoid repetition the complete assessment rubric for each item is presented only in Chapter 4, and only a general description of each of the SOLO levels is given for each task in the methodology section. Each assessment rubric identifies the SOLO level and the exemplars or criteria for each level.

The application of the large population sample model to the contextual task of the Mt. Wellington cable-car task provides an illustrative example of analysis used. A prestructural response described a student's response that was unable to use the sample size model and showed little development as a consequence of the study. A unistructural response described a student's response where the sample size model was partially applied, but the student was unable to interpret the result sensibly. A multistructural response described a student's response where the model was used to calculate the measurement accuracy as a formal mathematical task, but not all aspects were integrated. A relational response was demonstrated where the student attended effectively to all aspects of the task, demonstrated conceptual understanding of sample size, and used and interpreted the sample size model correctly. An extended abstract response enlarged a relational response to include relevant aspects not specifically included in the scope of the task.

Individual students' responses were coded as either PS, U, M, R, or EA for the levels of pre-structural, unistructural, multi-structural, relational and extended abstract responses. The number of students at a specific SOLO level was calculated as a percentage of the class group, and the relative proportion of students at each SOLO level was determined. Exemplars of students' work for each of the SOLO level for items assessed using SOLO were chosen for inclusion in the body of the thesis. The assessments were not checked independently, but the researcher's two supervisors checked the coding of the exemplars examined which is a representative stratified sample of between ten and forty percent of all student's responses for each item. Only a limited number of disagreements occurred and these were resolved readily through discussion between the supervisors and the researcher.

Students' development of understanding was examined by a comparison of student responses on the national and state election survey (Sections 3.3.2.4 & 3.3.5.8) and the Mt. Wellington cable-car (Sections 3.3.4.7 & 3.3.5.6).

3.8.4 Evaluation of items using instrumental genesis and situated abstraction

Instrumental genesis (Section 2.5.6.1) provided the analytical lens for the examination of how the Fathom software acted upon the user to support learning and students' development of use of the Fathom tool as a mathematical instrument. Central to instrumental genesis is the notion of schemes, which are the mental processes needed to use the software tool effectively (Section 2.5.6.1) To support examination of students' development of schemes this study borrowed from situated abstraction the three-step epistemic approach of Hershkowitz et al. (2001). The analysis was not intended to provide a comprehensive use of instrumental genesis framework, but rather to provide additional insights and evidence of students' learning trajectories.

As the use of instrumental genesis is relatively novel, an example is given from the study. In this study a student was presented with the artefact of a blank Fathom workspace and the student assembled a Fathom die simulation, but the Fathom die became an effective instrument only when the student accepted the virtual die as a fair representation of a die. When instrumental genesis is applied to the die simulation this allows the diagram presented earlier as Figure 2.3 to be refined and re-presented as Figure 3.3. Schemes are largely invisible and elusive, and their identification is somewhat problematic and subjective, but in this instance the two schemes associated with the development of the instrument are first, the procedural process of assembling the die simulation and second, the acceptance of the die as fair.

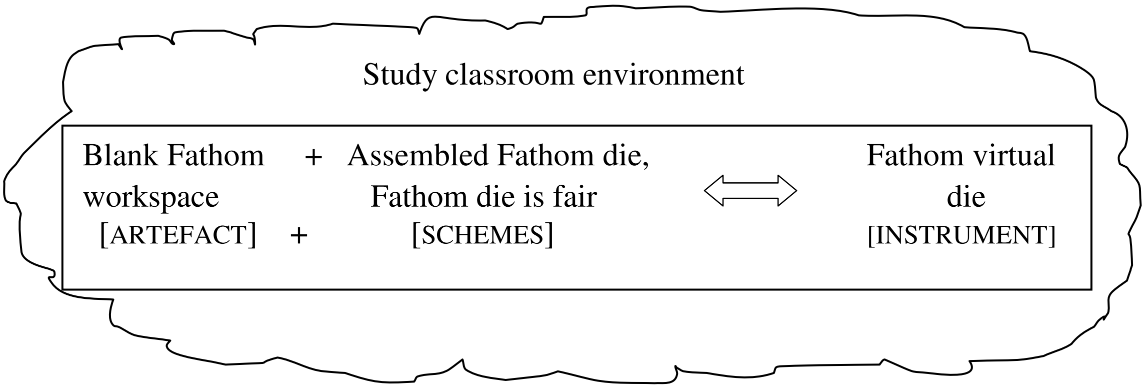


Figure 3.3. An example of instrumental genesis as applied in this study.

Of principal interest in this study is instrumentation, which is where the software acts upon the user. Students began the study with no knowledge of the Fathom die simulation; they could initially either consider the simulation as fair, unfair, or defer forming an opinion until having the opportunity to examine the simulation. The process of testing the Fathom die and comparing the random nature of its behaviour with other die provided the opportunity for the process of instrumentation and the software to act upon the student to occur. The evolutionary and personal nature of instrumental genesis seemed self-evident: students' development of use of the software and students' acceptance of the simulation (and the extent of that acceptance) occurs within an individual student's own time-frame and learning pathway.

The framework was used to identify specific instances of how the software appeared to support students' transition from (a) frequency to proportional data analysis, (b) language use of "tossing a coin" to "sample size," (c) small to large sample sizes, (d) interpreting a graph to choosing a graph to "tell a story" and (e) elementary to more sophisticated interpretation of a graphical representation of the cumulative proportion of heads.

The terminology of situated abstraction was incorporated into instrumental genesis to examine how students abstracted mathematical meaning. Hershkowitz et al. (2001) three step epistemic process of *constructing* new knowledge, *recognising* an existing mathematical structure as relevant, and *building-with* to apply knowledge to a new context provided by the terminology used. This structure was used to describe specific instances of students' mathematical learning of the fairness measure, students' development of understanding of the re-sampling terminology of "sample size used to calculate a measure" and "the number of measures collected," and students in the detailed study's interpretation of graphical representations of the cumulative proportions of heads.

Instrumental genesis was also used in this study to examine students' development of procedural use of Fathom. Procedural use was defined as the basic skills required to complete a Fathom task, such as construct a graph or summary, and procedural use of Fathom to construct graphs and summaries are examples of schemes. The software was used to support learning and to allow students to attend to the mathematical concepts – developing skills in the use of the software was incidental to the mathematics, but

Fathom could be a potential constraint on learning if students found the software difficult to use or it was not introduced effectively.

The study focussed on students' use of Fathom at four points in the classroom study at Lessons 1, 3, 5, and on the post-study assessment, and these four points provide a developmental path for analysis (Section 4.5.4.1). Lesson 1, the exploratory data analysis of the New York marathon data set, produced students' response to their first use of Fathom, and Lesson 3 gave students' response to their first use of a Fathom simulation. Lesson 5 compared male and female students' responses to two different versions of the one Fathom worksheet, and the post-study assessment assessed students' independent use of Fathom under traditional examination conditions. In all four instances students' responses were assessed as the proportion of students able to use Fathom effectively such that Fathom was not likely to be a barrier to learning. These data were supported further in the detailed study by an examination of students' ability to demonstrate their recall of the four basic operations of taking a sample, creating a summary, changing sample size, and creating a graph. In the detailed study students were assessed as the proportion of students able to complete "by self," "after demonstration," and "after demonstration and support."

3.8.5 Evaluation of post-study student questionnaires

The thirty questions presented to students were coded on a five-point Likert scale, and their responses were subsequently consolidated to the three responses: "strongly disagree" and "disagree" were coded as a negative response; "maybe" was coded as a neutral response; and "agree" and "strongly agree" were coded as a positive response. Students' responses were analysed as the proportion of students that provided each category of response. The responses of the boys and girls were then compared.

3.8.6 Colleague teacher interview transcription and analysis

Each interview was recorded to a separate audio file. The interviews were manually transcribed to a word processing document and the time elapsed was logged. The colleague teacher was identified by the letter "T" and the researcher was identified by the letter "R."

The interviews were analysed using an iterative process. In the first instance major themes were identified. These major themes were subsequently categorised as the student cohort, the benefits of the study, the appropriateness of the topics for Year 9, class discussion, students' acceptance of Fathom simulation, and Fathom software as a teaching and learning tool (Appendices C.2 & C.3). The teacher's responses were re-expressed as condensed meanings (Kvale, 2007), and the condensed meanings were re-presented as one table (Appendix C.4) categorised by the colleague teacher and the theme.

The data from the six themes were used to provide additional supporting evidence for students' acceptance of the Fathom simulation, the role of classroom discussion, the Fathom as a teaching and learning tool, and the appropriateness of the topics for Year 9 students.

3.8.7 Detailed study interview transcription and analysis

Each interview was recorded to a separate audio file. The interviews were manually transcribed to a word processing document and the time elapsed was logged. Individual students were identified by the first letter of the first name, and the researcher was identified by the letter "R".

The detailed study interviews were also analysed using an iterative process. In the first instance major themes were identified. These major themes were subsequently categorised and coded as nine themes of the interpretation of Graphical data representations (G), use of key Terminology including measures and sample size (T), expressed beliefs of Sample size (S), ease and accuracy of decimal Calculations (C), consideration of survey Accuracy (A), perceived belief of fairness of a Die (D), development of Proto-theories (P), use of Fathom as a mathematical tool (F), and any occurrence of students' off-task Behaviour (B) (Appendices D.2 – D.8). Students' responses were re-expressed as condensed meanings (Kvale, 2007), coded, and the condensed meanings were re-presented as one table (Appendix D.9) categorised by student pair and theme.

This comparison of students' responses was then used to support the study's first research question through examination of students' acceptance and preference for the Fathom die; and the third research question through students' use of re-sampling terminology; interpretation of numeric data representations to support students' transition from additive, to proportional, to distributional thinking; interpretation of measures dot plots and preferences for the cumulative proportions of heads graph; and use of Fathom.

3.9 Classroom work samples grouped by the three research questions

The classroom work samples and the detailed study tasks are grouped by each of questions and re-presented as Tables 3.8 – 3.10.

Students' development was established by comparing the level of students' responses on initial, developmental, and final tasks, supported by data from the detailed study. Several items provided data for more than one of the three research questions. For brevity Research question 1 is abbreviated to "Developing acceptance of the Fathom die simulation," Research question 2 is abbreviated to "Is $e = \pm 1/\sqrt{n}$ an accessible sample size model?" and Research question 3 is abbreviated to "Fathom re-sampling as a tool for high school."

Research question 1 was addressed by an examination of the development of students' beliefs of the fairness of a physical die through to the introduction and use of the Fathom virtual die. Table 3.8 presents the work samples that provided data for Research question 1.

Table 3.8.
Work Samples for Research Q.1: Developing Acceptance of the Fathom Die Simulation

Aspect of research question	Classroom study			Detailed study
	Initial task	Developmental task	Final task or assessment	
Students' beliefs of the die simulation	Physical die (Section 3.3.2.2) Data spread of a class set of a multiple coin toss (Section 3.3.2.3)	Home-made die (Section 3.3.3.1) Develop a fairness measure (Section 3.3.3.2) Fairness measure homework (Section 3.3.3.4) Fathom virtual die – first Fathom simulation (Section 3.3.3.5) Compare three dice using GICS (Section 3.3.3.6) Evaluation of fairness of Fathom die relative to a factory-made die (Section 3.3.3.6)	Post-study questionnaire items (Section 3.4 Items 7, 12, 14, 21, & 25)	Part D: Three potentially biased virtual dice (Section 3.7.5)

Research question 2 considered whether the sample size model was accessible to high school students. The three criteria used to select the large population sample size model (Section 2.4.12) were that the model was potentially accessible to students such that students could use and apply the model sensibly, that the model revealed underlying statistical concepts and promoted sound intuitive understanding of sample size and measurement accuracy, and that the use of the model recognised students' development lay on a continuum that was built on foundations in earlier years and supported more formal study at senior school levels (Section 2.4.12).

To address these criteria Research question 2 examined the five aspects of students' beliefs of sample size, use of models of the single statistic of the fairness measure, use of the sample model, interpretation of survey accuracy, and application of the sample size model in contextual tasks. Table 3.9 presents the work samples grouped by the five aspects.

Table 3.9.

Work Samples for Research Q. 2: Is e= ±1/√n an Accessible Sample Size Model?

Aspect of research question	Classroom study			Detailed study
	Initial task	Developmental task	Final assessment	
Beliefs of sample size	Sample size for a national and state election – Pre-test (Section 3.3.2.4)	The effect of sample size on the fairness measure (Section 3.3.4.1) Physical coin toss – cumulative proportion of heads (Section 3.3.4.4) Compare intuition of a 50 tosses with a Fathom coin toss (Section 3.3.4.6)	Sample size for a national and state election – follow-up test (Section 3.3.5.8). Post-study student questionnaire (Section 3.4 Items 10, 11, 19, 20 & 23)	Part H: For and Against – contextual sampling task (Section 3.7.9)
Models of the form of a single statistic	Develop a fairness measure (Section 3.3.3.2)		Fairness measure homework (Section 3.3.3.4) Coin measures 50 & 500 tosses of a coin homework – Part 2 (Section 3.3.4.3)	Part C: The effect of sample size on preference for data representation (Section 3.7.4)
Models of the form of a function $e = \pm 1/\surd n$	Sample size function (Section 3.3.2.1)	Large population sample size model $e = \pm 1/\surd n$ (Section 3.3.4.8)	Mathematics of the sample size model – post-study (Section 3.3.5.4) Post-study student questionnaire (Section 3.4 Items 13 & 17)	
Interpretation of survey accuracy	Data spread of a class set of a multiple coin toss (Section 3.3.2.3)	Fathom virtual 50 & 500 tosses of a coin simulation (Section 3.3.4.5) Compare intuition of a 50 tosses of a coin with a Fathom coin toss (Section 3.3.4.6)	Federal election survey: Howard and Rudd election survey (Section 3.3.5.2). Post-study questionnaire (Section 3.4 Items 26 & 28)	Part F: Cumulative proportion of heads graph and sample size (Section 3.7.7)
Use of large population sample size model in contextual tasks	Mt. Wellington cable-car – naive (Section 3.3.4.7)		Mt. Wellington cable-car – Post-study test. (Section 3.3.5.6) Post-study student questionnaire (Section 3.4 Item 23)	

Data to examine Research question 3 were collected by consideration of three aspects of students' development of peculiar to the use of Fathom and re-sampling. The first aspect recognised the importance of language in learning and considered the use the key terminology of sample size used to collect a measure and number of measures collected. The second aspect recognised the role of graphical representations of data and considered students' use of the novel representation of measures dot plots introduced for this study. The third aspect considered students' more general relationship with Fathom through students' development of use of Fathom in the classroom, how Fathom may have promoted learning, and students' perception of Fathom, Table 3.10 presents the work samples for each of the aspects.

Table 3.10.
Work Samples for Research Q.3: Fathom Re-sampling as a Tool for High School

Aspect of research question	Classroom study			Detailed study
	Initial task	Developmental task	Final task or assessment	
Key terminology	Develop the Fairness Measure (Section 3.3.3.2)	Coin measures homework – Part 1 (Section 3.3.4.2)	50 students in a Year 9 maths class (Section 3.3.5.1) Post-study student questionnaire (Section 3.4 Item 6)	Part F: Cumulative proportion of heads and sample size graph (Section 3.7.7)
Measures dot plot	Female race-times, Pre-test Q.5 (Section 3.3.2.1, Q. 5)	Compare three dice using GICS (Section 3.3.3.6) The effect of sample size on the fairness measure (Section 3.3.4.1). Coin measures homework – Part 2 (Section 3.3.4.3)	Mixed up measures (Post-study assessment Q. 3 (a) Section 3.3.5.3) Badly biased coin (Post-study assessment Q. 4 Section 3.3.5.5) Post-study student questionnaire (Section 3.4 Item 1 & 8)	Part F: Cumulative proportion of heads and sample size graph (Section 3.7.7) Part G: Measures dot plots at sample sizes of 50 & 500 (Section 3.7.8)

Table 3.10. (cont.)

Work Samples for Re. Q.3: Fathom Re-sampling as a Tool for High School (cont.)

Aspect of research question	Classroom study			Detailed study
	Initial task	Developmental task	Final task or assessment	
Students' relationship with Fathom	New York marathon – introduction to Fathom (Section 3.3.2.5)	Fathom virtual die – first Fathom simulation (Section 3.3.3.5) The effect of sample size on the fairness measure (Section 3.3.4.1)	Fathom basic skills test (Section 3.3.5.7) Post-study student questionnaire (Section 3.4 Item 3, 4, 18 & 29)	Part A: Die one face, or six faces presented simultaneously (Section 3.7.2) Part B: Coin side or both sides presented simultaneously (Section 3.7.3) Part C: The effect of sample size on preference for data representation (Section 3.7.4) Part D: Three potentially biased virtual coins (Section 3.7.5) Part E: Graphs and the Law of Large Numbers (Section 3.7.6)

3.10 Summary

This chapter describes the methodology of the research study. The research study was conducted as a three-week classroom teaching unit in two Year 9 classes in two government high schools. The classroom study was supported by pre and post-study testing of the students, a post study questionnaire, a detailed study of six pairs of students who had participated in the classroom study, and post-study interviews with the two colleague teachers.

The classroom study was conducted in four phases that broadly followed the three research questions. Phase 1 examined students' pre-study understanding of concepts and their beliefs of aspects of probability and sample size. Phase 2 sought to cultivate and assess students' acceptance of Fathom simulation through a process of statistical enquiry that examined and compared the fairness of a home-made die, a factory-made die, and a Fathom virtual die. Phase 3 of the study examined students' development of understanding of sample size and the use of the large population sample size model. Phase 4 provided the post-study formal assessment, which was conducted under traditional examination conditions.

The detailed study of six student pairs conducted approximately two months after the conclusion of the detailed study provided additional data to complement the classroom study. Five aspects were examined: (a) students' acceptance of the Fathom virtual die, (b) the use of key terminology of sample size used to calculate a measure and the number of measures collected, (c) students' interpretation of numeric data representations and how students might be supported in developing proportional and distributional thinking, (d) the use and interpretation of measures dot plots introduced in this study, and (e) the development of the procedural use of Fathom.

Students' development of understanding of concepts were examined using SOLO, students' development of use of Fathom was examined using principles derived from instrumental genesis, and students' development of understanding of specific mathematical aspects was examined using the terminology of situated abstraction. This thesis now turns to the data collected from both the classroom and the detailed study.

CHAPTER 4 Results

4.1 Introduction

This chapter examines the results collected from the data collection instruments used in the four phases of the classroom-based teaching sequence and the detailed student study. The three research questions provided the basis for the research study.

- **How effective is a statistics education research best-practice based approach of scientific enquiry in developing high school students' acceptance of the Fathom virtual simulator?**
- **In what ways does the sample size model $e = \pm 1/\sqrt{n}$ provide an accessible method for high school students to explicitly determine sample size when sampling from large and infinite populations?**
- **In what ways does this study's pedagogical approach of using Fathom virtual simulation and re-sampling offer an effective learning opportunity for high school students? What affordances and constraints do students encounter?**

The first section in this chapter provides an analysis of the pre-test that assessed students' basic mathematical skills and their ability to interpret a dot plot, and students' beliefs about a coin toss and of sample size when sampling from a large population (Section 4.2).

The second section provides an analysis of students' acceptance of the Fathom virtual simulation (Section 4.3). Students' acceptance of the simulation as a legitimate mathematical tool was traced through developmental tasks, the final assessment, the post-study questionnaire, with additional information provided by the detailed students study and the colleague teacher interviews. Students' pre-existing beliefs of the simulator were not assessed because students had no prior knowledge of the simulation tool.

The third section provides an analysis of the students' development of understanding of use of the sample size model $e = \pm 1/\sqrt{n}$ (Section 4.4). This was assessed through pre-testing, an initial task of sample size, developmental tasks, final evaluation tasks, and a test item designed to assess students' long-term recall of the model. Conceptual understanding of sample size and measurement error was also assessed.

The fourth and final section provides an analysis of students' development and use of re-sampling techniques using Fathom (Section 4.5). Specifically this section examines three aspects peculiar to re-sampling and Fathom: key terminology associated with re-sampling, interpretation of measures dot plots, and students' relationship with Fathom that includes .procedural use of Fathom, instances where Fathom supported learning, and attitudes to Fathom.

4.2 Pre-study assessment

4.2.1 Basic mathematical skills (Pre-test Q. 1 (a–f) & Q. 5)

The objectives of these items were to contribute to the development a student profile of basic mathematical skills thought essential for the study: fractions and percentages (Q1); substitution of an integer into functions of the form of the sample size model (Q. 1(e)); and interpretation of a basic graph presented in an unfamiliar format (Q. 5). The methodology for the tasks is provided in Section 3.3.2.1, the items are attached as Appendix A.1, Q.1, and the analysis of the data is presented in Appendix G.2. The data for Q. 5 are subsequently re-examined in more detail in Section 4.5.3.1 to provide data for Research question 3.

The male class had little difficulty with calculating fractions and percentage, but only 42% (14/33) of the female class, answered all four items correctly, a result that was unexpected for students enrolled in a mathematics extended class. The students had little difficulty in interpreting correctly the measures dot plot of female race-times.

4.2.2 Physical die (Pre-test Q. 4)

The objective of the item was to determine students' ability to assess the fairness of a die based on the column graph of a die rolled thirty times. The item varied the work of Watson and Moritz (2003) by presenting students with the data of a die trial, rather than the physical die. Methodology for the task is provided in Section 3.3.2.2, and the item is attached as Appendix A.1, Q. 4.

Table 4.1 shows that the students' default position was to assume the die was fair, and that the die's behaviour was attributable to chance. Three students (two boys and one girl) were prepared to consider the die was possibly unfair, and all three proposed rolling the die before reaching a conclusion. The potential effect of the way the die was rolled or other physical conditions was mentioned by both male (9.5%) and female (15.1%) students. Only one student responded that the data lay outside the expected behaviour of the die and the die was unfair.

Table 4.1.
Students' Pre-study Responses to Whether the Data of a 30 Roll of a Die Showed that the Die was Fair

Response	Male students		Female students	
	No.	%	No.	%
Die fair or unfair, no justification	0	0.0 %	1	3.0 %
Die fair, rolling method used	2	9.5 %	5	15.1 %
Die unfair, outside normal variation	0	0.0 %	1	3.0 %
Die fair, within normal variation	16	76.2 %	19	57.6 %
Die is possibly unfair, requires larger sample size or insufficient evidence	2	9.5 %	1	3.0 %
No response	1	4.8 %	6	18.3 %
Total	21	100.0 %	33	100.0 %

Three students thought the die could not be anything other than fair, and several made assumptions or included evidence that was not provided in the question.

> No, because all die faces are equally balanced [Student S1001J]
>
> …a die is unable to be biased [Student N2610H]

One student argued that the fact that one face had occurred more frequently demonstrated that the die's behaviour was random.

> The fact that 3 occurred more often that the other numbers showed that the results were truly random. [Student X0211G]

Other students recognised that a different outcome would have occurred if the die were rolled again, but they did not consider changing the sample size.

> …tossed another day another number would possibly appear most [Student I0711G]
>
> No, if you tried it again it would probably be different. [Student S2503C]

If students compared the frequency, the comparison was made to the face that occurred the next most frequently (face 5 occurred six times), rather than to the expected frequency.

> …it [Face 3] was only thrown three times [more] than 5 [Face 5]. [Student N3110T]

Student S0403J provided an example of a student who would increase the sample size before making a decision.

> I would only agree [that the die was unfair] if more tests were conducted. [Student S0403J]

4.2.3 Data spread of class set of a multiple coin toss (Pre-test Q. 2 & 3)

The objectives of the two items were to establish students' pre-existing intuition of the range where *most* of a class set of the proportion of heads from fifty tosses of a coin would occur. The term most was used in this study to establish students' personal natural language equivalent for a 95% confidence interval. The item was re-presented to the students subsequently as the 50 & 500 coin toss (Section 4.4.7) to allow students to compare their intuitions with data generated by a Fathom simulation. Methodology for the task is provided in Section 3.3.2.3, and the task is in Appendix A.1, Q. 2 & 3.

The first item asked students to select from five alternatives the range of proportions of heads that would occur when a class of 30 students tossed a coin 50 times (Table 4.2). Male and female student groups provided two distinctly different sets of responses. The most common response by the male students (6/21, 28.6%) was the proportion of heads 0.5 +/- 0.05 [i.e., 0.45 to 0.55] (Table 4.2, Q.2 c). This was a much narrower distribution than most commonly chosen by the female students (36.4%), who favoured the responses 0.5 +/- 0.1 [i.e., 0.4 to 0.6] (Table 4.2, Q.2 a). The boys' preference for a narrow distribution was consistent with a belief in small samples or it demonstrated the student strategy of choosing the middle of the multiple-choice alternatives offered. The female students appeared to be either more cautious, or had prior knowledge, or simply chose the broadest range because logically most results must occur within the broadest range provided. The most common female response was also the preferred response.

A proportion of male (5/21, 23.8%) and female (12/33, 36.4%) students did not provide a response. These students were either unfamiliar with the task or they were unfamiliar with +/- notation. Given that the item was multiple-choice and the confidence students demonstrated subsequently it is surprising that students did not at least attempt the task.

Table 4.2.

Evaluation of Students' Pre-test Item Q.2 The Distribution of Proportion of Heads where "Most" of a Class Set of a 50 Tosses of a Coin Occurred

Coin toss distribution		Male students		Female students	
		No.	%	No.	%
(a) 0.5 +/- 0.1*	(i.e. 0.4 – 0.6)	2	9.5 %	12	36.4 %
(b) 0.5 +/- 0.07	(i.e. 0.43 – 0.57)	3	14.3 %	5	15.2 %
(c) 0.5 +/- 0.05	(i.e. 0.45 – 0.55)	6	28.6 %	2	6.1 %
(d) 0.5 +/- 0.02	(i.e. 0.48 – 0.52)	3	14.3 %	1	3.0 %
(e) 0.5 +/- 0.01	(i.e. 0.49 – 0.51)	2	9.5 %	3	9.1 %
No response		5	23.8 %	10	30.3 %
Total		21	100.0 %	33	100.0 %

* The preferred response and the closest alternative to a 95% confidence interval

The second part of the task asked students to provide a definition for their use of the term "most," i.e., "what was your own personal definition of 'most' students (in the class of 30 students)?" The term most was proposed by the researcher as a natural language definition of a 95% confidence interval, and the question is clearly open to interpretation. Table 4.3 shows that both male and female students considered most students as a proportion of the class equal to or smaller than 25 of 30 students. The term most was replaced subsequently in the study by the stronger term "almost all."

Table 4.3.

Evaluation of Students' Use of the Term "Most" as a Natural Language Equivalent of a 95% Confidence Interval

Coin toss distribution	Male students		Female students	
	No.	%	No.	%
(a) 20 students in a class of 30	6	29 %	8	24 %
(b) 25 students in a class of 30	12	57 %	18	55 %
(c) 28 students in a class of 30*	1	5 %	1	3 %
(d) 29 students in a class of 30	1	5 %	1	3 %
No response	1	5 %	5	15 %
Total	21	100 %	33	100 %

* Closest to 95% confidence interval

Individual student's personal definitions of most, e.g., "28 students in a class of 30" were then paired with the range that the individual student specified, e.g., "(c) 0.5 +/- 0.05" (i.e., 0.45 – 0.55). Good estimates were considered either the pair of [(b) 0.5 ± 0.07 and (a) 20 students in a class of 30] or the pair [(a) 0.5 ± 0.1 and (c) 28 students in a class of 30]. Table 4.4 shows that a substantial proportion of the male students

(61.9%) predicted the data would occur in a narrower distribution than could be expected theoretically by chance – a response consistent with a belief in small samples. No male student provided a correct response. The female students were more cautious and accurate, but their responses differed little from a randomly chosen response. All the students who provided the good estimate gave the second of these two pairs of responses: [(a) 0.5 ± 0.1 and (c) 28 students in a class of 30]. Approximately one third of both male and female students did not provide a response to both questions.

Table 4.4.
Evaluation of Students' Intuitive Sense of the Distribution of the Proportion of Heads of a 50 Tosses of a Coins

Students' Intuitive Sense of a Coin Toss	Male students		Female students	
	No.	%	No.	%
Narrower distribution than can be expected theoretically by chance	13	61.9 %	9	27.3 %
Good estimate*	0	0.0 %	6	18.2 %
Wider distribution than can be expected theoretically by chance	2	9.5 %	7	21.2 %
No, or incomplete, response	6	28.6 %	11	33.3 %
Total	21	100.0 %	33	100.0 %

* Good estimates are considered either the pair of [(b) 0.5 ± 0.07 and (a) 20 students in a class of 30] or the pair [(a) 0.5 ± 0.1 and (c) 28 students in a class of 30].

4.2.4 Sample size for a national and state election survey (Pre-test Q. 6 & 7)

The objectives of the two items were to establish students' naïve understanding of sample size when sampling from a large population, to identify the strategies students use when choosing a sample size, and to explore whether students consider population size a factor when choosing a sample size. The contexts were public opinion surveys held immediately prior to national and state elections, with the national and state elections providing the two different population sizes. The item was a multiple-choice: students selected a sample size from the four multi-choice alternatives (see Table 4.5, a–d), chose a sampling strategy "that best described their thoughts" also from four alternatives (see Table 4.6, a–d), and students could include additional comments if they wished. Methodology for the item is presented in Section 3.3.2.4, the item and the tasks are attached in Appendix A.1, Q. 6 & 7.

In the item on the national survey, presented in Table 4.5, both male (62%) and female (55%) students favoured a sample size strategy of 10% of the population. Where students chose a specific numerical sample size their choice of sample size broadly decreased with sample size. The two strategies, which are identified as either a percentage or a numeric strategy, suggest students strongly preferred a large sample size and a sample size that is considerably larger than is used conventionally.

Table 4.5.
Students' Responses to Pre-test Q. 6 Sample Size for an Opinion Survey Prior to an Australian National Election

Sample size strategy	Male students		Female students	
	No.	%	No.	%
(a) About 10% of the population	13	62 %	18	55 %
(b) 15,000	6	29 %	6	18 %
(c) 1,500*	2	9 %	7	21 %
(d) 150	0	0 %	0	0 %
No response	0	0 %	2	6 %
Total	21	100 %	33	100 %

* Note: The accepted sample size when sampling from large populations

In the companion item on the smaller population state election opinion survey (Table 4.6) both male and female students demonstrated a substantial shift away from a percentage strategy, and a shift from larger to smaller numeric sample sizes. The proportion of males who nominated the preferred sample size of (c) 1500 increased from 9% to 33%, but the proportion of females who gave this response was unchanged. The proportion of both male and female students who did not give a response increased. The data suggest that students preferred a smaller sample size for a smaller population, but perhaps students did not recognise that a sample size of 10% of a smaller population would yield a smaller sample size.

Table 4.6.
Students' Responses to Pre-test Q. 7 Sample Size for an Opinion Survey Prior to State Election

Sample size strategy	Male students		Female students	
	No.	%	No.	%
(a) About 10% of the population	4	19 %	13	40 %
(b) 15,000	4	19 %	3	9 %
(c) 1,500*	7	33 %	7	21 %
(d) 150	2	10 %	5	15 %
No response	4	19 %	5	15 %
Total	21	100 %	33	100 %

* Note: The accepted sample size when sampling from large populations

Table 4.7 considers whether students adopted consistent or inconsistent sample size strategies for both the national and state surveys. Students predominately adopted a consistent strategy: a sample size strategy of 10% of the population was used for both national and state survey by almost half (45.5%) of the girls and slightly under one third (28.6%) of the boys, and a consistent numeric strategy by one third (33.3%) of boys and 18.2% of the girls. Two students, one boy and one girl, gave the preferred response to both items.

Table 4.7.

Students' Sample Size Strategies for Opinion Surveys Prior to National and State Elections: "What best describes your thoughts?"

Sample size strategy	Male students		Female students	
	No.	%	No.	%
Inconsistent strategy, combination of 10% of population and a numeric.	6	28.6 %	5	15.1 %
Consistent "10% of the population" for both national and state polls.	6	28.6 %	15	45.5 %
Consistent numeric sample size with a smaller sample for the smaller state population.	7	33.2 %	6	18.2 %
Consistent numeric strategy of 15,000 for national and state election.	0	0.0 %	0	0.0 %
Consistent numeric sample size of 1500 for national and state election.*	1	4.8 %	1	3.0 %
Incomplete or no response.	1	4.8 %	6	18.2 %
Total	21	100.0 %	33	100.0 %

* Preferred response

The second part of items Q. 6 and 7 asked that students choose, from the multiple choice alternatives offered, the strategy that most accurately reflected their thoughts in choosing a sample size. Table 4.8 shows that similar proportions of boys (29%) and girls (30%) gave "practicalities and cost" as the principal consideration. Almost a quarter of the boys (24%), but only one girl, volunteered a separate explanation to the effect of "use as large a sample as possible to improve a survey's accuracy." Over half of the girls (52%), and 14 % of the boys, gave the explanation of "eliminated a few and the guessed," which could be interpreted as the students reaching for a solution or recognising a strategy used in tests widely. It was an acknowledgement that students had little background knowledge of sample size. One student only gave the response "knew it from school," and this confirms that the topic of sample size was not considered extensively at school. Of the one boy and one girl who gave the preferred sample size of 1500 only the male identified the media as the source of information.

Table 4.8.
Students' Explanations for the Source of Their Sample Size Strategy: "What best describes your thoughts?"

Response	Boys		Girls	
	No.	%	No.	%
(a) Knew it from newspapers or TV	1	5 %	1	3 %
(b) Considered practicalities and cost	6	29 %	10	30 %
(c) Eliminated a few and guessed	3	14 %	17	52 %
(d) Knew it from school	1	5 %	0	0 %
Other: accuracy, take largest sample	5	24 %	1	3 %
Other: intuition	2	9 %	0	0 %
Other: unclassified	0	0 %	1	3 %
No response	1	14 %	3	9 %
Total	21	100 %	33	100 %

A sampling strategy of 10% of the population was used extensively, with over half of the students preferring this strategy. Students were mindful of the cost and practicalities of conducting a survey. The boys also considered, prior to the tuition of this study, the accuracy of a survey. The sampling strategies and the explanations for the strategies suggest students were attempting informal sense-making of sample size concepts.

4.2.5 New York marathon – introduction to Fathom

The objective of this activity was to introduce students to the software Fathom. An exploratory data analysis task of a New York marathon race times was used – rather than a probability simulation – because this was thought contextually more familiar. This was a peer-tutored activity where students worked in pairs; one student who had received prior instruction in the software by the researcher explained the software to the other student in the pair. To complete the activity successfully students developed sufficient familiarity with the basic software, used their existing skills of data analysis to interpret the data, and prepared a written response to a question of their own choosing. The task gave a benchmark of students' abilities to use Fathom after their first use of the software. Methodology for this task is provided in Section 3.3.2.5, the task is attached as Appendix A.2, and scanned copies of class sets of Fathom workspaces are attached as Appendices F.2 and F.3.

In the class discussion students proposed questions such as "which runner had the most wins?" ;"which country had the most wins?"; "did the times change over the years?" An extract from the researcher's journal, verified by the colleague teacher subsequently at the colleague teacher interview, noted that students did not have any difficulty using Fathom with the minimum of tuition. Students produced a range of responses from unsophisticated to moderately sophisticated that showed high engagement with the task. The peer-tutored activity, requiring ten minutes of instruction by the researcher to one student in the pair, proved a effective and time-efficient method to introduce Fathom to students. This result is consistent with earlier studies using Fathom (Bill, 2007).

Table 4.9 shows a large difference between the performance of the male and female students, but the explanation is clear: the professional journal (June 25, 2008) noted that there was insufficient time for the female students to complete the task. Of the male students who submitted a work sample 62% gave a response assessed as multistructural or higher.

Table 4.9.
SOLO Evaluation of New York Marathon Times Data Analysis Task

SOLO level	Male students		Female students		Student exemplar
	No.	%	No.	%	
P	1	7 %	0	0 %	– Incomplete
U	3	21 %	11	61 %	– Grete Waitz won
M	6	34 %	7	39 %	– Slowest men's time similar to fastest women's time
R	2	14 %	0	0 %	– Female times didn't drop dramatically like the male times
EA	2	14 %	0	0 %	– Comparisons are difficult given changing training and diet
Total	14	100%	18	100%	

Student E1709S provided a unistructural response that presented a simple description of the graph without providing any data analysis.

> Which countries have the most and least wins and how long it took them to complete the race. [Student E1709S]

Student S1808B provided a multistructural response. The student foreshadowed use of the GICS framework by providing a global view of the data – for this student the GICS framework formalised what the students did naturally. The student integrated several aspects of the data, e.g., race times, number of winners, and country of origin, but the analysis contained several errors of analysis, e.g., the average race-time for combined male and female times was nonsensical for the two distinct cohorts and the numeric data were presented to an inappropriate number of decimal places. The students would not have had the knowledge of Fathom to calculate separate means for male and females and the unfamiliar units of decimal hours may have distracted the student.

> Above I have graphed the race-times of the NY marathon. The average race-time, for both male and females, is 2.9305 hours. There are 30 male winners and 30 female winners in the data. USA have the most wins overall with Grete Waitz of Denmark (female) being the individual to win the most races, winning nine times. The fastest time is 2.13361 hours, run by Jurna (male) of Tanzania, with the slowest time being 3.14472 hours run by Nina Kuscsik (female) of USA.. [Student S1808B]

Student N2103S provided a relational response. The student identified significant data points, calculated correctly the difference between male and female race-times, and considered the effect of the year on both male and female race-times.

> In 1970 when the race was started [first occasion] was the slowest male winner [….] The slowest female winner was in 1971. As the years passed the races got faster [….] in 1989 a Tanzania ran the fastest time […] so far no-one has beaten his time. [….] the difference between the fastest female and male time is: 2.42139 h – 2.13361 h = 0.28778 h difference. It took more years for the female time to decrease than it did for the male time. [Student N2103S]

Student S1610J provided an extended abstract response that identified the behaviour of a trend (related two variables), compared the male and female winners, and included information from outside the task. The student was engaged with the data set, but the student showed some minor misunderstanding of the marathon.

> There was a distinct trend in the data which shows that the earlier the marathon was held, the slower the winning time. It drops rapidly before bottoming out at about 2.2 hours for males and 2.45 for females [….] when the times appear to flatten out it could be due to the participants reaching their highest possible skills level. [….] this model would only be accurate while attributes remain the same, including course distance, medical technology, and other environmental attributes. [Student S1610J]

A common feature of students' responses was a Fathom workspace cluttered with a number of data representations that did not connect strongly to their analyses, or graphs of such complexity that the underlying information was not displayed effectively. Other students were distracted by cosmetic or superficial features such as colour. This first Fathom task was both an exploratory data analysis and a software exploratory task, and students used the opportunity to constructively explore the features of the software, for example, one student attempted to use the linear regression model feature.

The activity provides the first opportunity for analysis using the instrumental genesis framework (Section 2.5.6.2). The students were presented with the bare artefact of a Fathom worksheet containing the collection of the New York marathon race-times. Students used pre-existing schemes, for example, existing knowledge of the data set, data quoted to an appropriate number of significant places, interpretation of graphical data representation, in combination with newly constructed schemes, for example, constructing a graph in Fathom through a sequence of dragging and dropping a graph icon and attributes, choosing a graphical representation and attributes appropriate for the question posed, using the comments box to include a verbal analysis, and managing the Fathom worksheet workspace effectively.

Students demonstrated a range of sophistication of newly constructed schemes. Students needed to integrate the new schemes associated with Fathom and their existing knowledge of data analysis. Student S1610J described the graph using informal language of “drops rapidly,” “bottoming out,” and “flattens out.” In other instance Fathom acted to confound learning; students S1808B and N2203S quoted race-times to an inappropriate five decimal places. These examples of instrumentations where the software acted upon the student could support learning, but the software could also act against learning.

4.2.6 Summary of findings

This sub-section sought to establish students’ pre-study beliefs and basic mathematical competencies thought essential for the study. The basic mathematical skills test showed that the male students possessed the skills thought essential for the study, but that a substantial proportion of the female students found arithmetic tasks involving percentages and fractions difficult. Three-quarters of the male students and two-thirds of the female students performed a basic calculation using the reciprocal of the square correctly (Pre-test Q. 1 (e)). When assessing the data of 30 rolls of a physical die students’ default positions were predominately that the die was fair, and fewer than 10% of both male and female students thought that the die was possibly unfair and a larger sample size was needed. When considering the distribution of the proportion of heads of a series of multiple coin tosses male students predicted a narrower distribution than can be expected by chance, and the female students’ responses differed little from that of a randomly chosen response. When choosing a sample size for an opinion survey prior to the large populations of national and state elections both male and female students favoured a sample size of 10% of the population, which far exceeds the sample size used conventionally, and there was some evidence that students tended to favour a smaller sample size when surveying a smaller population. This suggests that students' knowledge of sample size when sampling from large populations was modest. Students were introduced to Fathom through a peer-tutored exploratory data analysis task, and students had little apparent difficulty using the basic Fathom features productively within the first lesson.

4.3 Research Q.1: Developing acceptance of the Fathom die simulator

4.3.1 Introduction

This sub-section examines students' development of acceptance of the fairness of the Fathom die using a process of statistical enquiry that investigated physical and virtual dice. Students examined random behaviour through an investigation of the fairness of three dice: a Home-made die students fabricated using modelling clay, a conventional Factory-made die, and a Fathom virtual die. A formal statistic, the Fairness Measure, developed for the study was used to compare the three dice. Students' acceptance of the Fathom virtual was established through post-study questionnaire items and in the detailed study. The methodologies for these tasks are presented in Section 3.3.3, and the work samples are summarised in Table 4.10.

Table 4.10.
Work Samples for Research Question 1: Developing Acceptance of Fathom Virtual Simulation

Classroom study			Detailed study
Initial task	Developmental task	Final task or assessment	
Physical die (Section 4.2.2) Data spread of a class set of a multiple coin toss (Section 4.2.3)	Home-made die (Section 4.3.2) Develop a fairness measure (Section 4.3.3) Fairness measure homework (Section 4.3.4) Fathom virtual die – first Fathom simulation (Section 4.3.5) Compare three dice using GICS (Section 4.3.6) Evaluation of fairness of Fathom die relative to a factory-made die (Section 4.3.6)	Post-study questionnaire items (Section 4.3.7 Items 7, 12, 14, 21, 25)	Part D: Three potentially biased virtual dice (Section 4.3.8)

4.3.2 Home-made die

The objective of the activity was to introduce simulation using a physical die, but this familiar activity was varied to the conventional approach and students fabricated their own die using Sculpey™ modelling clay. This potentially unfair die provided stimulus for students to consider the random behaviour of a die. Students rolled the die 30 times and recorded the frequency at which each face occurred. Methodology for the task is provided in Section 3.3.3.1 and the worksheet is attached as Appendix A.3.

Several students, perhaps thinking the researcher expected a fair die to be created, felt they had done so. The comment also hints at the subsequent strategy as a simulation "fair enough for our purposes."

> For a home-made die I think it is as fair as you could get it. [Student E2205J]

Students considered the physical appearance or symmetry of the die as part of their assessment of the die's fairness. Students were legitimately arguing from the available evidence; if that were the extent of the analysis then that would be an unsophisticated response, but it could be part of a sophisticated response. Students used a range of terms including cube, rectangle, and even.

> The die isn't technically a cube. The table shows it favours 3 but it is pretty fair. [Student E2810J]
>
> No [not fair], it is an uneven rectangle so it is hard for it to land on the 3 and 4. [Student H1112I]
>
> For a home-made die, it is quite fair. However, not every side is even which could affect the accuracy of the result. [Student I0812A]

Students were clearly aware of the tension between examination for bias, and the role of chance, and were at least prepared to suspend their judgment. Several students argued strongly from the evidence, and the language adopted suggests a semi-formal statistical approach.

> From our data we were not convinced that the die was fair or unfair because we believed it was chance. [Student E2611G]
>
> The home-made die is sort of fair as each side was hit a similar amount of times. It could be said to be unfair as the number 4 was landed on more frequently. [Student R1207L]
>
> I don't think the die is totally fair, however I don't think we can be sure either way [….] should have been on each side [face occurring] from 3-7 or 4-6, but in actual fact the range was 2-8 [….] but in the end it is chance, but it could be biased. [Student R1610A]

Sample size was explicitly considered by three students only. All three considered the sample size of 30 too small, and student H1306E considered a sample size of 200 appropriate as part of a strategy to determine the fairness of a die. Student E1709S, mindful of the die's shape, expected the die to be unfair. The criterion for fairness of the dice was developing from informal sense to formal approach.

> I think the die was pretty fair, although the five seemed to appear more frequent than the other numbers. If you kept rolling the die I think the five would get more in front of other numbers. [Student L2007E]
>
> The die seems to be unfair but I don't think this can be determined with only 30 rolls. [Student E1709S]
>
> Roll the die more times and see if a certain number comes up more than the rest, say 200 times. [Student H1306E]

Manifestly unfair dice or extreme results allowed students to make a definitive, almost deterministic, judgment. Such extreme results potentially obscured the concepts related to the random behaviour of rolled dice, but this was an inherent risk of using simulation activities in the classroom.

> No [not fair], because the die seems to be biased towards one and six. [occurred 11 and 10 times respectively] [Student N0106N]

4.3.3 Develop a fairness measure

Students were asked to develop a formal objective measure of the fairness of dice: "We need one number that measures how unfair the die is." Methodology for the task is provided in Section 3.3.3.2, and the item is attached as Appendix A.3, Q.5.

Students provided one of three categories of response (a) no response, (b) a descriptive verbal, but not necessarily formal mathematical, approach, and (c) a mathematical approach oriented towards calculating a single statistic. A verbal description is a developing mathematical response where students cannot fully translate to a formal mathematical language.

Table 4.11 shows that almost half of the female students did not give a response; the girls were either not engaged with the task or not sufficiently confident to give a response. Only 15% of females attempted a formal mathematical response. The boys' responses were strikingly different to the girls' responses with 94% of boys providing a descriptive verbal or attempting a formal mathematical approach. All of the exemplars presented here were considered multistructural because the students attempted a formal mathematical approach.

Table 4.11.

Students' Development of a Single Statistic to Measure the Fairness of a Die

Students' responses	Male students No.	%	Female students No.	%
No response	1	6 %	15	47 %
Descriptive (Unistructural)	7	44 %	12	38 %
Attempts formal mathematical approach (Multistructural)	8	50 %	5	15 %
Total	16	100 %	32	100 %

Several students considered the difference between their observed and expected frequencies, or compared the most and least frequently occurring faces; these approaches are suggestive of the statistic used ultimately, and given sufficient time it is conceivable the students could have developed the statistic independently. Two of the proposals also included criteria for acceptance or rejection of fairness. Several students proposed a statistic similar to that used in the study, which suggests that the statistic was accessible to the students.

> [The] difference between highest and lowest is not more than 4. [Student I0711G]
>
> Look at the average difference between our results for each face and what we expected. [Student S1610J]
>
> Any single number [that] gets above 7 [a frequency of seven] is unfair. [Student Y0305R]
>
> Maybe we should get the total and divide by the amount of times it came up per number [face]. For example, six came up once so there is, based on this a 1/30 chance. [Student G0709A]
>
> [The] deviation from the 1/6th of sample average. [Student N2610L]

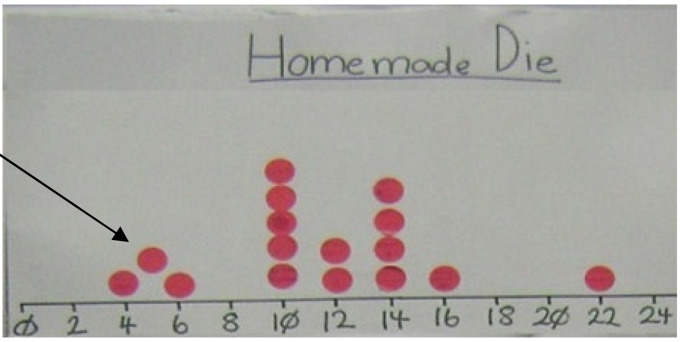
Two students proposed a formal statistic of the fairness of a dice identical to that used in the study. It was not clear whether or not the two students developed the idea independently of each other.

> Find difference between observed and expected frequencies for each face and then add them all together. [Student N2306C]

One student, R1610A, provided a very sophisticated response proposing repeating the test with a factory-made die and comparing the data with that of the home-made die. The student created a statistic of the average of the class results. Such an approach lays the foundation for statistical techniques comparing distributions against a standard.

> Try again on a factory made die and compare results. Get an average of class results. [Student R1610A]

A whole-class discussion explored students' proposals for the fairness measure (Section 3.3.3.3), and the discussion concluded with researcher suggesting that a fairness measure calculated as the sum difference between observed and expected be used (See Section 3.3.3.2). A calculation of this fairness measure was demonstrated by the researcher, and students used the worksheet to calculate the fairness measure (Appendix A.3) for their own dice. Students calculated the fairness measure for the home-made die without any difficulty, and added their one data point to the class display of the fairness measures (Figure 4.1); the fairness measure of five was calculated incorrectly. Students conducted the same test with the Factory-made die, rolled the die 30 times, calculated the fairness measure, and included the data in a dot plot measures graph of the Factory-made die (See Figure 4.5 subsequently).



Student incorrectly calculates fairness measure of five.

Figure 4.1. Fairness measure dot plot of home-made die displayed as a class poster.

Students, within the one activity demonstrated all three elements of situated abstraction (Section 2.5.6.2). They took their mathematical knowledge of the expected and observed behaviour of the die and elements of simple arithmetic to construct, analyse, and ultimately build-with to develop a mathematical structure of a proto-type fairness measure. The nested behaviour of situated abstraction seemed apparent with all three elements intertwined. The students then used the same procedure to test the factory-made die, calculated the fairness measure, and recorded the fairness measure on a separate dot plot (See Figure 4.5).

4.3.4 Fairness measure homework

The objectives of the task were to promote mathematical rigour and understanding by exploring the mathematics of the fairness measure, to provide opportunities to practice sub-skills, to provide a topic for researcher-led discussion and act as a final assessment of the mathematics within the fairness measure. The methodology is presented in Section 3.3.3.4, and the worksheet is attached as Appendix A.5.

All students who submitted responses (84% of girls and 62% of boys) had little difficulty in completing all four elements of the task. The most difficult task for students was Q. 3, which asked students to reverse the conventional order of analysis and provide a graph based a given fairness measure, rather than the more familiar task of creating a graph from the raw data of a coin toss. Two girls and two boys did not complete this element of the homework. Students M1306E and N2306C produced two unconventional graphs (Figures 4.2 & 4.3). When the item was analysed using the language of situated abstraction students built-with their existing knowledge of the fairness measure to apply that knowledge to develop and demonstrate a more thorough understanding of the mathematics within the fairness measure. This activity suggested students who submitted responses had a sound understanding of the mathematics of the fairness measure, and that the fairness measure was within the grasp of the students.

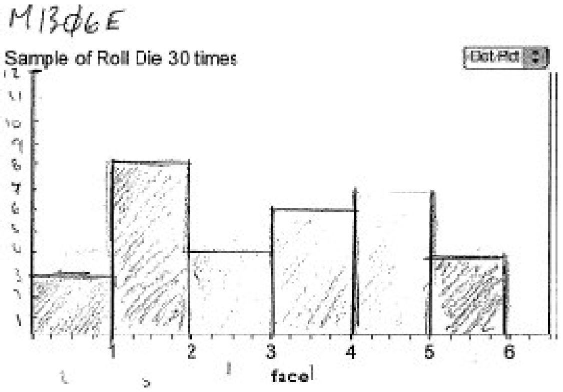


Figure 4.2. Student M1306E did not centre columns on tick mark.

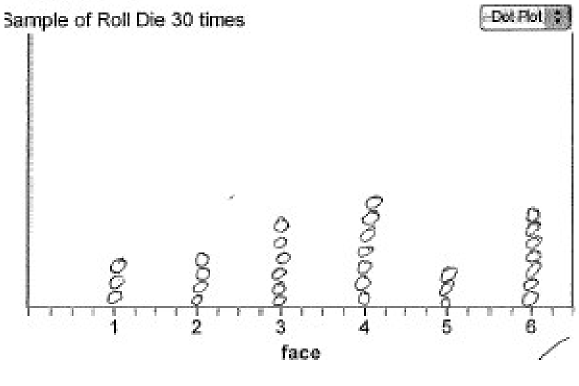
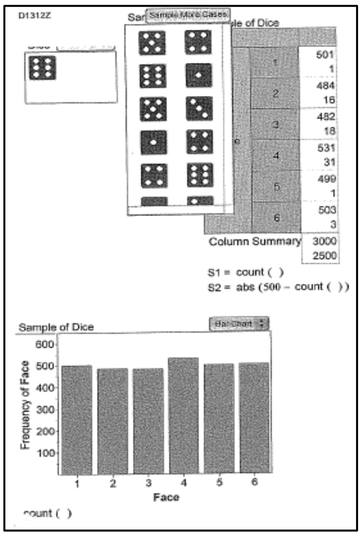


Figure 4.3. Student N2306C preferred a dot plot representation.

4.3.5 Fathom virtual die – first Fathom simulation

The objective of the virtual Fathom activity was for students to assemble and test the Fathom die using the same test procedure that was used with the two physical – the home-made and the factory-made – dice. Methodology for the activity is provided in Section 3.3.3.5, the worksheet is attached in Appendix A.6, and scanned copies of class sets of students' Fathom workspaces are attached as Appendices F.4 and F.5.

Students had no apparent difficulty in assembling the Fathom die using the guided worksheet and all students produced (but not necessarily submitted) a functioning simulation. Students' workspace worksheet management had improved from the New York marathon exploratory data analysis task (Section 4.2.5) with the workspaces clear and uncluttered. Several students explored the simulation independently by changing the sample size from the default 10, to larger sample sizes such as 600 or 3000. Two students also used the formula editor independently to calculate the difference between the observed and expected frequencies using the abs (absolute value) function, which is a step towards calculating the fairness measure (Figure 4.4).



1	501 1
2	484 16
3	482 18
4	531 31
5	499 1
6	503 3
Column Summary	3000 2500

S1 = count ( )
S2 = abs (500 – count ( ))

Formula editor used on student's own initiative

Figure 4.4. Student D1312Z Coin simulation and use of formula editor.

4.3.6 Compare three dice using GICS

The objectives of this task were to evaluate students' ability to compare three data distributions of the fairness measures using the Global, Individual, Measures of Centre and Measures of Spread (GICS) framework, and to examine students' analysis of whether they considered the Fathom die less fair, fair as, or fairer than the standard factory-made die. During this task students worked independently under traditional examination conditions, but students could use notes taken during the class discussion and the classroom poster of the fairness measure was displayed prominently at the front of the class (Figure 4.5). Given the length of student responses, exemplars for each of the three SOLO levels are attached in Appendix F.1, and only summaries and extracts of the exemplars are presented here. The methodology for the task is presented in Section 3.3.3.6, and the worksheet is attached in Appendix A.7.

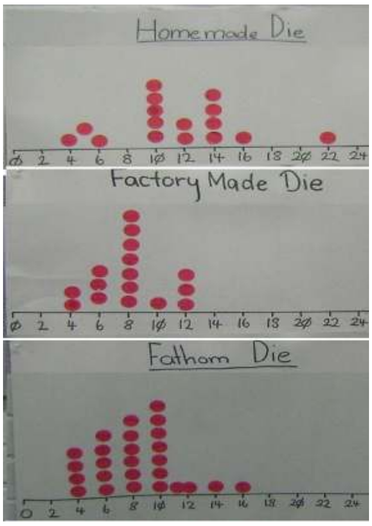


Figure 4.5. Fairness measure dot plot of home-made, factory-made, and Fathom dice used in the girls' school.

Table 4.12 shows that the levels of the girls' responses were high, and almost all the girls (97%) provided a multistructural response or above. The standard of the boys' responses was of a far lower level than the standard of the girls' responses.

Table 4.12.

SOLO Evaluation of Students' Analysis of the Fairness of Three Dice Using the GICS framework

SOLO level	Male students		Female students		Exemplars or Criteria
	No.	%	No.	%	
NR	4	19 %	0	0 %	No response
U	5	24 %	1	3 %	Only single aspects are considered, and in a disconnected fashion.
M	9	43 %	19	59 %	Considers all GICS four aspects, but all four aspects are not integrated fully.
R	3	14 %	12	38 %	Considers all four GICS aspects in an integrated way.
Total	21	100 %	32	100 %	

In the second part of the item students compared the fairness of the Fathom die to the standard reference of the factory-made die and provided a justification that the Fathom die was less fair, as fair as, or fairer than the factory-made die. Table 4.13 shows that more male students felt that the Fathom die was fairer than the factory-made die, and more female students believed the Fathom die was less fair than a factory-made die.

Table 4.13.
Evaluation of Students' Acceptance of the Fathom Die as "Fair" Relative to a Conventional Factory-made Die

Perceived fairness of Fathom relative to a factory-made die	Male students		Female students	
	No. of students	%	No. of students	%
No response	5	24 %	1	3 %
Less fair	3	14 %	19	59 %
As fair	4	19 %	9	28 %
Fairer	9	43 %	3	10 %
Total	21	100 %	32	100 %

The analysis then considered if students justified their belief of the fairness of the die on the available data or they used some other criteria; that is, whether or not the students' opinions were evidence-based. Table 4.14 shows that about 2/3rds of both male and female students expressed an opinion based on the evidence, a higher proportion of the female students (8/32, 25%) than the male students (3/21, 14%) did not explicitly use the data to form an opinion, and 19% of male students did not provide a response.

Table 4.14.
Evaluation of Criteria Students Used to Determine Whether or Not the Fathom Die was "Fair" Relative to a Conventional Factory-made Die

Perceived fairness of Fathom relative to a factory-made die	Male students		Female students	
	No. of students	%	No. of students	%
Argues principally from the available evidence	14	67 %	23	71 %
Response not evidence based	3	14 %	8	25 %
No response	4	19 %	1	4 %
Total	21	100 %	32	100 %

That the Fathom die was less fair than the factory-made standard die was argued from either scepticism of the technology or rationally from the available data. Only three students remained suspicious of the technology. These three students did not argue strongly from the available evidence, and thought that the lack of transparency of the simulation was a significant factor.

> Because you can never trust technology and the computer might even out the occurrences of each number. [Student S0212M]
>
> Because you don't know how the computer is calculating the results and you can never trust technology. [….] however the Fathom die is quite fair. [Student I0812A]
>
> I would say less fair because it is a computer doing the working out where as you know what you've rolled if you are doing it yourself. [Student N3993A]

The female students' group argued legitimately on the available evidence that the Fathom die was "less fair" than the factory-made die, because the data suggested that the Fathom die was "less fair." The explanation lies with way the data were collected The research journal (July 5th, 2008) recorded the colleague teacher's observation that students continued to run the simulation until a more interesting, either unusual or extreme, result was obtained. This was an inherent risk of using simulation as a teaching tool – the simulation might not necessarily produce an output that directly supports the concepts being taught. Students E2810J, E1709S and R1610A provided exemplars of students arguing from the available data.

> The dot plots on the wall indicate that the fathom die goes up to 16 on the fairness scale while the factory-made die goes up to 12 on the fairness level. [Student E2810J]
>
> Because the Fathom made die has dots on 14 and 16 whereas the factory made die stops at 12, although it is normal for a die to have a sudden peak at 14 or 16. [Student E1709S]
>
> Because the spread is larger and the peak [centre] is higher. [Student R1610A]

Students who believed the Fathom die was "as fair" as a factory-made die demonstrated an awareness that the fairness differed, but they did not consider this difference as meaningful. For example some students noted that these differences were not consider significant or were attributed to chance. The combined use of informal terms and the correct and incorrect use of formal statistical terms indicated students' developing use of statistical language.

> [….] because both were pretty fair and I believe that were both due to chance. [Student M0306E]

> The Fathom die software is unbiased and the observed frequency of face numbers is random or based on chance. [Student R2408I]
>
> Although there was a slight difference on the average, the Fathom spreads was lower down. [Student E3011L]
>
> The factory and Fathom-made die are more evenly distributed [….] Home-made results are more scattered. [Student N2610H]
>
> The factory made and the Fathom die have the same fairness as they have very similar distributions and any apparent bias can be attributed to random variation. [Student S1610J]

Three girls and nine boys argued that the Fathom die was “fairer” than the factory-made die. To make this claim students examined the data and included in their analysis a consideration of whether a person can influence the outcome of a die. Two students noted the apparent lack of human involvement in running the Fathom simulation, although one student stated that a computer cannot be biased.

> [….] human error doesn’t affect it at all. Chance is the only thing playing a part. [Student R0308I]
>
> … because a computer cannot be biased. A computer has no opinion whereas a person can purposefully roll a die the way he/she wants. [Student X0211G]

A third student, T1705A, argued entirely from an interpretation of the data.

> [….] the Fathom die had a more evenly distributed chance of falling on each face judging by the dot graph result. [Student T1705A]

The fairness of the Fathom die was investigated as a statistical enquiry, and an essential element of statistical enquiry, purposefully cultivated during the study, was that conclusions must be evidence based. Of the students who submitted responses 62% (13/21) of the male students and 87.5% (28/32) of female students argued their position from the available data.

Students who considered the Fathom die “as fair as” and “fairer than” the factory-made die (62% of males and 37.4% of females) demonstrated confidence in the simulation data. For these students, any lack of confidence in the die simulation appeared not to be a barrier to learning.

4.3.7 Students’ post-study questionnaire items

At the conclusion of the classroom study students’ beliefs in the randomness and confidence in Fathom virtual simulation were examined in the post-study questionnaire in Items 7, 12, 14, 21, and 25, the questionnaire is attached as Appendix A.18, and the data for these items are presented in Table 4.15. Questions are paraphrased for brevity.

That students accepted the Fathom die as fair was a demanding criterion and not essential for the study: the study sought only that students' perception of fairness and legitimacy of Fathom were not a barrier to learning. Students' acceptance of the simulator was high with 83.3% (15/18) of the male and 88.6% (31/35) of the female students either agreeing or strongly agreeing with Item 7 "The Fathom die was as random as it needed to be for what we were doing." One student only disagreed, and the remaining students were neutral.

Students' confidence in the Fathom die may be influenced by whether the die behaved as the students expected, given their experiences using a physical die. This was explored in Item 12 "I was surprised how many times I had to roll the Fathom die before I thought it was random;" 22.2% (4/18) of male and 17.1% (6/35) of female students were surprised at the number of times the Fathom die was rolled before they accepted the die was random. In response to Item 25, "I didn't have confidence in Fathom because the results were too weird" 16.7% (3/18) of males and 8.6% (3/35) of females thought the data generated by Fathom were peculiar. If students had largely accepted that the Fathom simulation was fair, reservations regarding the random behaviour of the simulation remained amongst some students.

Item 14, "Building the simulation myself gave me confidence in the simulation," considered if assembling the virtual die had contributed to students' acceptance of the simulator. Of the male students 72.2% (13/18) and 54.3% (19/35) of females thought that assembling and step-wise checking of the simulation were important in developing their confidence. The assembling and checking process may be important to developing acceptance amongst students generally, but the female students may have been more likely to accept the simulation on trust than the male students.

Students' acceptance of the Fathom die relative to the familiar physical factory-made die was explored through Item 21, "I had more confidence in a physical die than a Fathom die." At the conclusion of the classroom study, and despite specifically attending to beliefs, at least one third of both male and female had more confidence in the physical die than the Fathom virtual die.

Table 4.15.

Post-study Questionnaire Students' Acceptance of the Fathom Die

Post-study questionnaire item		Disagree or Strongly disagree %	Maybe or neutral %	Agree or Strongly agree %
Students' perceived fairness of the Fathom die				
7. By the end of the unit I was convinced the Fathom coin was "as random as it needed to be for the work we were doing"	M	5.6%	11.1%	83.3%
	F	0.0%	11.4%	88.6%
Fathom die behaved as students expected				
12. I was surprised how many times I had to "roll" the Fathom die before I thought the die was random	M	83.3%	0.0%	22.2%
	F	65.8%	17.1%	17.1%
25. I didn't have confidence in the Fathom die because some results were too "weird"	M	77.8%	5.5%	16.7%
	F	77.1%	14.3%	8.6%
Assembling and step-wise testing of the Fathom die developed confidence				
14. Building the simulation myself gave me confidence in the simulation	M	5.6%	22.2%	72.2%
	F	11.4%	34.3%	54.3%
Confidence in the Fathom die relative to confidence in a standard factory-made die				
21. I have more confidence in a physical die than a Fathom die	M	38.9%	27.8%	33.3%
	F	25.7%	37.1%	37.1%

4.3.8 Detailed study workshop

Students in the detailed study were asked their preference for the physical or the Fathom virtual die. The physical die was intuitively more real, more fair or random than the virtual die, and students rarely hesitated when expressing their preference for a physical die.

> Physical. [die and coin] [Student R1706D]
>
> I'd probably go for the physical die. Real or not it's the one that I have a gut-instinct preference. [Student N2610H]
>
> The factory-made die. [Students S1001J & T0612M, simultaneously and without hesitation]
>
> I think the factory-made is more fair, but I'd rather use computer [because] it is quicker. [Student S1001J]
>
> Always thought it [Fathom] wasn't completely random. [Student Y1504L]

Students were aware of the efficiency and speed of the Fathom simulation. When given a choice between a physical and virtual die the advantage of using a computer at large sample size when a physical die was impracticable was clear: students did not hesitate in preferring the Fathom die at large sample sizes. The virtual coin and die had the disadvantage that the simulation had to be assembled and that the software was not available outside of school. The advantages of the physical die and coin included the portability and ready availability of a die and coin. Arguably this was intelligent use of the available tools. At a more subtle level, and recalling the class discussion, two students noted the virtual die and coin eliminated the bias potentially introduced in using a physical die.

> With a computer die you can do can do a lot more, thousands, it's quicker, I wouldn't use a real [physical] die. [in that situation] [Student Y1504L]
>
> The Fathom coin does take out the physical biases. [Student R1706D]
>
> [….] but then again you can take the random elements out of a physical die as well [by] throwing it a certain way. [Student Y1504L]
>
> If it was only ten rolls I use the real [physical] one. [….] If I had the option of computer, the computer, but it is harder to carry around. [Student N2701B]

One student thought that the additional difficulty of assembling a Fathom simulation was an impediment, but once assembled and operating the simulation was quick. Assembling the simulation was one of the schemes of the instrumental genesis.

> It was easier to do things physically because you weren't as good in getting set up in Fathom, but much quicker in rolling the die in Fathom. [Student E2611G]

Part D explored the information – the sample size – students required when they made a decision regarding the fairness of a virtual die. The sample size is clearly related to the level of the bias evident. For subtly biased dice with expected values between 0.48 and 0.53 students were prepared to make a decision on whether the coin was biased using sample sizes between 40 to 140 – sample sizes not dissimilar to a physical simulation. One student pair made a decision on bias at a sample size of 420 (Appendix D.8).

4.3.9 Colleague teacher interview

Both colleague teachers felt that, by the end of the program, the students had accepted the Fathom virtual simulation as legitimate. Both teachers noted the advantage, for students and teachers, of the speed and ease with which the data could be generated by the simulation.

> The students accepted the simulation, no longer questioned the tool, and they had faith in its ability to model a reality. The [Fathom] die was more than a pretend die and they tried to transfer the information to other situations. [….] Collecting data from physical systems is very time-consuming and difficult. This time-constraint is eliminated once students developed a belief and a confidence in the random behaviour of the simulator. [Colleague teacher of male class]

> I didn't hear any mumblings or dispute about the Fathom die and coin – they seemed to accept it. They would have used the [conventional physical] die in Years 7 and 8. The students recognised that it made modelling more efficient. [Colleague teacher of female class]

Asked whether the elaborate process of developing confidence in the random behaviour of the Fathom simulation by comparing the random behaviour of home-made, the factory-made and the Fathom die was effective, one colleague teacher thought that students found the activity fabricating and testing their own home-made die as particularly engaging, and the fairness measure dot plot as an effective technique to support the analysis of the data but the activity might be necessary only for younger or lower-ability students. A key decision for the teacher is to select appropriate material for any class. The colleague teacher the importance of students "trying it out," but one of the study's intentions' was to extend an informal approach to a more formal disciplined scientific approach.

> It depends on the class. A less able class would need the engagement of making the die, and this class liked it, but it was not essential. The students certainly needed the opportunity to "try it [the dice] out." [Colleague teacher of female class]

4.3.10 Summary of findings for Research question 1

This sub-section examined the first of the three research questions: whether the process of statistical enquiry was effective in promoting acceptance of the Fathom die simulation as legitimate for the students. Students were given with the opportunity to explore Fathom simulation through a process of statistical enquiry and mathematical experiences of substance including a formal statistic developed for the study of the fairness measure. Students compared the fairness of three dice: a home-made die fabricated using modelling clay, a conventional factory-made die, and the Fathom die simulation. The legitimacy of the Fathom simulation was investigated in terms thought comprehensible to the students as the Fathom die was "fair enough for our purposes."

Evidence for students' acceptance of the Fathom die was provided by the activity Compare three dice using GICS, items taken from the post-study student questionnaire, the detailed study interviews, and the colleague teacher interviews. Students' initial beliefs of the random behaviour could not be established readily; students would simply not know whether the Fathom die was fair, but by the conclusion of the classroom study students generally appeared to accept the Fathom simulation as "fair enough for our purposes." This evidence suggests that the process of statistical enquiry in developing students' acceptance of the Fathom die simulation as "fair enough for our purposes" was effective.

The instrumental genesis framework was used to introduce and use Fathom and to provide a means of analysing students' response to the software. Schemes are the mental processes needed to use the software tool for the task at hand. One example of a scheme is acceptance of the simulation as fair..

A second example of a scheme was that students had begun to internalise the basic procedural use of Fathom, and they were developing the procedural knowledge to use Fathom. Students were introduced to Fathom through a peer-tutored exploratory data analysis and introduced to Fathom simulation using a guided worksheet.. To use Fathom effectively students needed to manage the workspace; be familiar with use of terminology peculiar to Fathom of collection, attribute, and re-sample; take a sample, and create a graph and a summary. This is represented diagrammatically in Figure 4.6 by extending Figure 2.3 to include Fathom as the artefact, the two schemes of acceptance of the die as fair and procedural use of Fathom, to form the instrument of a Fathom probability simulator.

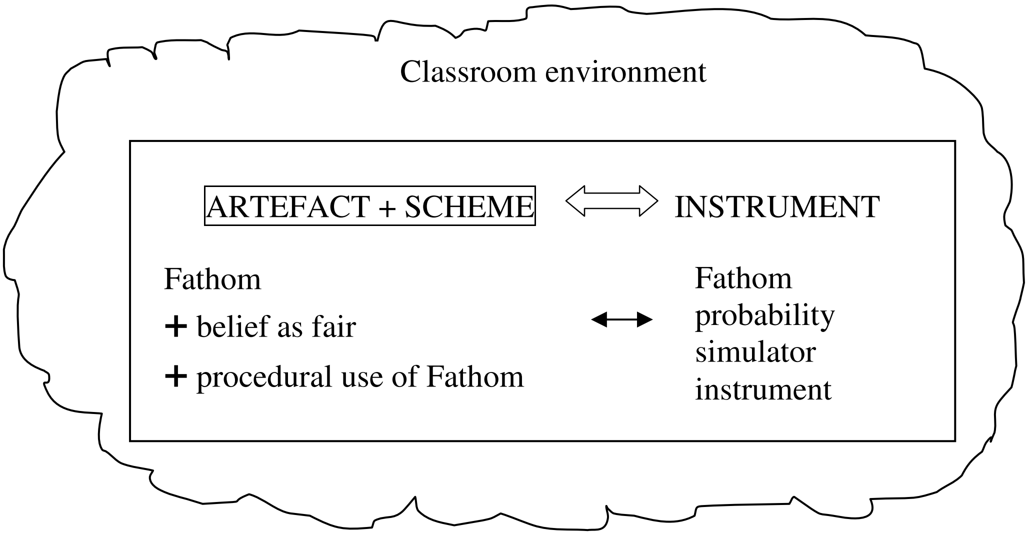


Figure 4.6. Instrumental genesis of a Fathom probability simulator.

The two schemes of beliefs and procedural use of Fathom have in common that neither proved a significant barrier to learning. The two schemes were interrelated because the students, particularly the males, thought assembling the simulation was important in their acceptance of the simulation. Instrumentation, where the software acted upon the user, may not always act to support learning, but Fathom did not appear to be a barrier to learning. The process of instrumental genesis was underway, and Fathom as probability simulation instrument was forming.

4.4 Research Q2: Is $e = \pm 1/\sqrt{n}$ an accessible sample size model?

4.4.1 Introduction

This section addresses the second of the three research questions: students' use of the large population sample size model $e = \pm 1/\sqrt{n}$ to explicitly determine sample size. Students' use of the sample size model was within a study that used Fathom simulation and sought to cultivate students' sense of sample size and accuracy of sampling.

Students' use of the model and their development of intuitive understandings were examined from the perspective of students' change in beliefs of sample size, use of models of the form of a single statistic, use of the large population sample model, interpretation of survey accuracy, and application of the sample size model in contextual tasks. Students' work samples, which were grouped by these five aspects and presented first in Table 3.8 ,are re-presented in Table 4.16.The large population sample size model was introduced and examined, in common with students' earlier examination of fairness of the three dice, through a process of statistical enquiry that sought to establish the sample size model's accuracy and utility. Students' ability to use the functions of the same form of the sample model in an elementary way was established in the pre-test by substituting values into the function, and the sample model was used subsequently in a series of formal mathematical and contextual public opinion surveys tasks. Students' consideration of sample size concepts and of the sample size model were evaluated by items from the students' post-study questionnaire, the colleague teachers' interviews, the detailed study, and two items taken from the previous sub-section examining the fairness of dice.

Students' ability to apply the large population sample size in the long term outside of the classroom was established through a follow-up test item conducted two months after the conclusion of the classroom study.

Table 4.16.
Work Samples for Research Question 2: Explicitly Quantifying Sample Size

Aspect of research question	Classroom study			Detailed study
	Initial task	Developmental task	Final task or assessment	
Beliefs of sample size	Sample size for a national and state election – Pre-test (Section 4.2.4)	The effect of sample size on the fairness measure (Section 4.4.3) Physical coin toss – Law of Large Numbers (Section 4.4.4) Compare intuitive sense of 50 tosses of a coin with a Fathom coin toss (Section 4.4.7)	National and state election – follow-up test (Section 4.4.13 & 4.4.14). Post-study student questionnaire (Section 4.4.15 Items 10, 11, 19, 20 & 23)	Part H: For & Against – contextual sampling task (Section 4.4.17)
Models of the form of a single statistic	Develop a fairness measure (Section 4.3.3)	The effect of sample size on the fairness measure, Q. 6 (Section 4.4.3)	Fairness measure homework (Section 4.3.4) Coin measures 50 & 500 tosses of a coin homework – Part 2 (Section 4.4.5)	Part C: The effect of sample size on preference for data representation (Sec. 4.5.4.2)
Models of the form of a function $e = \pm 1/\sqrt{n}$	Sample size function (Section 4.4.2)	Large population sample size model $e = \pm 1/\sqrt{n}$ (Section 4.4.9)	Mathematics of the sample size model – post-study (Section 4.4.11) Post-study student questionnaire (Section 4.4.13 Items 13 & 17)	
Interpretation of survey accuracy	Data spread of a class set of a multiple coin toss (Section 4.2.3)	Fathom virtual 50 & 500 tosses of a coin simulation (Section 4.4.6) Compare intuitive sense of 50 tosses of a coin with a Fathom coin toss (Section 4.4.7)	Federal election survey: Howard and Rudd election survey (Section 4.4.10). Post-study student questionnaire (Section 4.4.13 Items 26 & 28)	Part F: Cumulative proportion of heads graph and sample size (Section 4.4.16)
Use of large population sample size model in contextual tasks	Mt. Wellington cable-car – naive (Section 4.4.8)		Mt. Wellington cable-car – post-study test Q9. (Section 4.4.13 & 4.4.14)	

4.4.2 Students' background knowledge of the sample size function

The model introduced in the study, $e = \pm 1/\sqrt{n}$, was complex. The pre-test asked students to substitute the integer 9 into the expression $1/\sqrt{n}$ and complete the calculation manually. Methodology for the item is presented in Section 3.3.2.1, and the item is attached as Appendix A.1, Q.1 (h). The item is presented in full, but only the data for the third and final question, calculating the reciprocal of the square root function, relates to the sample size model and is of interest.

(h) If X = 9 $\sqrt{X}$ = X^2 = $1/\sqrt{X}$ =

Table 4.17 shows that 76% of male students and 64% of female students gave the correct response. The students who gave an incorrect response understood the square root symbol, but ignored the reciprocal and gave the answer as "3" rather than "1/3." Almost 20% of both male and female students did not give a response to this item, but calculated correctly the companion question √9. This suggests that many students were confounded by the compound operation of fractions, square root, and reciprocal.

Table 4.17.
Students' Responses to Pre-test Item Q.1(h) Calculation of Reciprocal Square Root Function

Student's	Male students		Female students	
response	No.	%	No.	%
Correct	16	76 %	21	64 %
Incorrect	1	5 %	6	18 %
No response	4	19 %	6	18 %
Total	21	100 %	33	100 %

4.4.3 The effect of sample size on the fairness measure

The principal objective of this task was to provide a transitional activity between the fairness measure of the die simulation and subsequent sample size activities involving coins by exploring the effect of sample size on the fairness measure. The fairness measure was re-calculated as the %fairness measure to allow comparisons to be made at sample sizes used of 30, 300, and 3000 rolls of a Fathom die. Methodology for the task is presented in Section 3.3.4.1, the worksheet is attached as Appendix A.8, and the results are presented in Table 4.18. The item is re-examined in Section 4.5.3.3 as a companion task to the coin measures 50 & 500 tosses of a coin homework item.

Figure 4.7 provides an illustrative example of the %fairness measure dot plots that students constructed in this whole-class activity, and which formed the basis of the class discussion and their analysis. The students used the Fathom simulation and contributed one %fairness measure for each of the three sample sizes of 30, 300, and 3000, located a measure of centre for each of the three sample sizes, and interpolated to sample sizes of 900 and 1600. This was a complex task for students. Table 4.18 shows that students' performance on this task was modest and their understanding of the concepts as developing only.

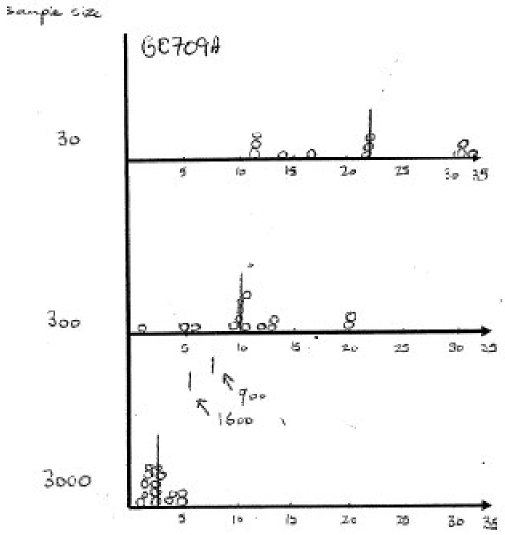


Figure 4.7. Student G0709A given as part of a relational response.

Table 4.18.
SOLO Evaluation of the Effect of Sample Size on the %Fairness Measure

SOLO level	No.	%	Q1, 2, 3, 4 & 6 male students only Exemplars or Criteria
P	3	16.7 %	Does not meaningfully attempt task
U	5	27.8 %	Demonstrates uncoordinated knowledge of single aspects of the task
M	7	38.8 %	Demonstrates a partially integrated, but not necessarily correct, understanding of the three aspects of the task
R	3	16.7 %	Integrates all three aspects of the task – Provides a meaningful graph correctly locating centres of each three distribution, labels axes, and includes a scale – Describes meaningfully the effect of sample size on centre and spread – Interpolates to other sample sizes and global describes the effect of an order magnitude change in sample size
Total	18	100.0%	

An example of a unistructural response included clearly constructed graphs marked with appropriately placed measures of centre and meaningful interpolations to sample sizes of 900 and 1600. The student's response was ambiguous, and the student appeared confused by the effect of sample size on the measure of centre with the actual location of the measure of centre, but the effect of sample size on the spread of the data was clear. The student's response consisted of specific uncoordinated elements. The student did respond to all questions.

> [How is the measure of centre affected by sample size?] The centre is staying with the biggest bunch of dots. [How is the spread affected?] It is getting smaller and more compact. [Student T0612M]

A multistructural response considered the effect of sample size on the spread of the measures, but not the location of the centre of the data distributions. The student incorrectly attributed the effect of sample size on the measure of centre to the use of a fairness measure based on percentage, but correctly noted the effect of sample size on the data spread.

> As the sample size gets bigger the variations aren't so much an effect because we are using the percentages. As the sample size gets bigger the spread gets less and less. [Is the die fairer at large sample size?] The effect of natural variation makes less of a difference [....] because we are dealing with percentages, not a plain number like 8 on the earlier thing [fairness measure] [Student D1312Z]

Student S1808B provided a relational response. The student identified a pattern, used informal, but appropriate, language, and the student combined existing knowledge and new knowledge.

> The lower the sample size [used] the higher the fairness measure [....] Once again the lower the sample size generally the range [of the data] is bigger. [Is the die fairer at large sample size?] The die doesn't change [in fairness]. The reason for the lower fairness measure is because it evens itself out. Also each number counts for less. [Does increasing the sample size 10 times reduce the fairness measure to a 1/10th?] No. Rolling the die is random so you can never be sure [....] It seems to decrease by a quarter. [Student S1808B]

One part of the item, Q. 6 explored students' understanding of the mathematical relationship between sample size and the %fairness measure, and the question was purposefully expressed to prompt a mathematical response, and asked, "Does increasing the sample size 10 times reduce the %fairness to 1/10th of its previous value?" The correct response is no, increasing the sample size ten times reduces the fairness measure by a quarter or third. Two broad categories of response were identified. The first was where students interpreted literally, confounded randomness with certainty, and did not provide a mathematical response. The second interpreted the task as the sample size having a tendency to reduce the %fairness measure. An example of the first category of response, where the response was unistructural because the student had taken the singular concept that the process was random, is provided by the following:

> No, it is random and there is a possibility it could be the same. [Student N2103S]

An example of the second, a correct and multistructural response because students demonstrated a distributional sense of the data, is provided by the following:

> No, it reduces it by roughly a quarter or a third. [Student G0709A]

Another example of second category of response was thought a multistructural response because the student considered both the random behaviour of the fairness measure and the effect of sample size. It is provided by this exemplar:

> No. Rolling a die is random, so you can never be sure of what the fairness will be. It seems to decrease by approximately a quarter. [Student S1808B]

Disentangling the effect of random natural variation behaviour and any bias that may exist was challenging for students.. Students who could appreciate the distinction between the two showed a significantly higher level of understanding of random phenomenon.

4.4.4 Physical coin toss (The Law of Large Numbers)

The task used a physical coin toss tossed 50 times to introduce students to a Fathom virtual coin simulation. The task included, at selected points within the 50 tosses of a coin, a calculation of “the difference between observed and expected” as a foundation for the subsequent introduction of the margin of error. Methodology for the task is presented in Section 3.3.4.4, and the worksheet is attached as Appendix A.11.

Students had little difficulty completing the task. Seventeen boys and 33 girls completed the calculations on the worksheet, but only 12 boys (71%) and 27 girls (82%) provided a response to the question “Did the value of the ‘difference from expected’ tend to get smaller, larger, or stay the same as the sample size increased?” Students’ responses ranged from single word responses, ambiguous responses, to more thoughtful analyses of the data. Examples of these responses included the following.

> Tended to vary evenly. [Student S0412N]
>
> Smaller. [Student N2103S]
>
> It generally decreased in size. [Student S1808B]
>
> Ours got smaller but near the end we had a big tail run so it got a bit bigger. [Student G1610I]

4.4.5 Coin measures 50 & 500 tosses of a coin homework – Part 2

The objective of the task was to assess students' ability to transfer their knowledge from the activity examining the effect of sample size on the fairness measure of a die (Section 4.4.3) to another, but familiar, context of the proportion of heads from a multiple coin toss. The students were given the item as homework, so the students worked largely independently and the item was given prior to using the Fathom coin simulation in class. Students were provided with a dot plot of the proportion of heads of a series of 50 tosses of a coin, and used that dot plot as a template to sketch the proportion of heads that would occur from a series of a larger sample size of 500 tosses of the coin. Methodology for the task is provided in Section 3.3.4.3, the worksheet is attached as Appendix A.10, Q. 2, and the results are presented in Table 4.19.

Only ten boys (48%) and 21 girls (64%) submitted a response. Although the data were unrepresentative of the class, the value of the task lay in gaining some insights into students' understanding before using the Fathom coin simulation. To complete the task successfully students constructed a graphical representation of the proportion of heads that included the features of a correctly located the centre of the data, a similar number of measures of proportion of heads and a narrower distribution as the 50 tosses of a coin measures dot plot.

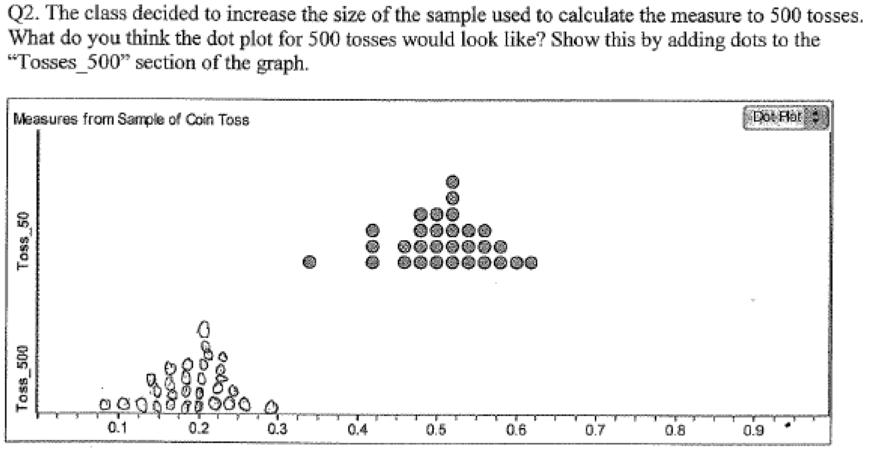
Table 4.19 this shows that of the students who submitted a response 70% of male students and 76% of female students provided a multistructural response. All students, when constructing the measures dot plot, provided the appropriate number of approximately 30 measures, and they did not confound the sample size with the number of measures.

Table 4.19.
SOLO Evaluation of Students' Sketching a 500 Tosses of a Coin Measures Dot Plot using a 50 Tosses of a Coins Template

SOLO level	Male No.	Male %	Female No.	Female %	Criteria or exemplars
U	1	10 %	1	5 %	Limited understanding with one or two elements of the criteria only.
M	7	70 %	16	76 %	Partial understanding, three criteria met.
R	2	20 %	4	19 %	Demonstrates complete and integrated understanding of concepts – Centre of 500 sample size dot plot located correctly – Similar number of measures for each dot plot – Distribution has a narrower spread than a 50 tosses of a coin
Total	10	100 %	21	100 %	

Students sketched the 500 tosses dot plot, but many students incorrectly displaced the centre of the distribution to the left (Figure 4.8) – the distribution for a fair coin is centred at a proportion of heads of 0.5. This was a common error: 11 of the 21 female students (52.4%) and two of the ten male students (20%) provided this response. These students appear to have misapplied the effect of sample size on the %fairness measure (refer Section 4.4.3 and Figure 4.6), which was displaced to the left as the sample size increased.

Student R1207L provided a unistructural response. The student sketched the 500 tosses dot plot, but displaced the centre of the distribution to the left (Figure 4.8). The student provided an appropriate number of measures, but the spread was only subtly narrower than for the smaller sample size of 50.



Q2. The class decided to increase the size of the sample used to calculate the measure to 500 tosses. What do you think the dot plot for 500 tosses would look like? Show this by adding dots to the "Tosses_500" section of the graph.

Figure 4.8. Student R1207L, proportion of heads sample size of 500.

Student I0812A provided a multistructural response. The student sketched a distribution of the proportion of heads for the 500 tosses of a coin toss narrower than for the 50 tosses of a coin, but the distribution was also centred to the left incorrectly (Figure 4.9). Again, this might be explained by the student taking a cue from "the effect of sample size on the fairness measure" activity shown in Figure 4.7.

Q2. The class decided to increase the size of the sample used to calculate the measure to 500 tosses. What do you think the dot plot for 500 tosses would look like? Show this by adding dots to the "Tosses_500" section of the graph.

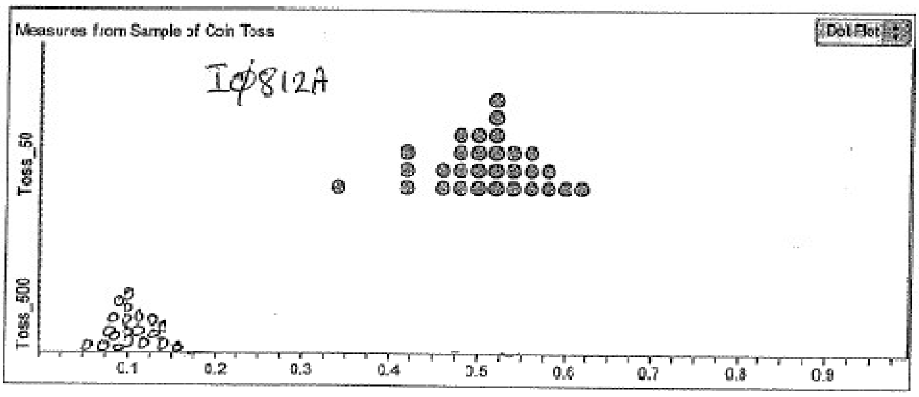


Figure 4.9. Student I0812A, proportion of heads sample size of 500.

Student N23006C provided a relational response that demonstrated an understanding of the concepts. The dot plot of the measures at the larger sample size of 500 tosses of a coin was placed centred correctly at a proportion of heads of 0.5 with a narrower spread of values and at least similar number of measures collected as for the 50 tosses (Figure 4.10).

Q2. The class decided to increase the size of the sample used to calculate the measure to 500 tosses. What do you think the dot plot for 500 tosses would look like? Show this by adding dots to the "Tosses_500" section of the graph.

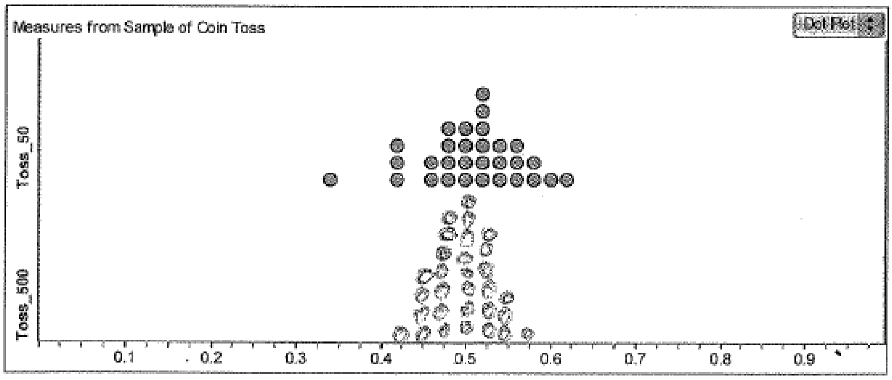


Figure 4.10. Student N2306C, proportion of heads sample size of 500.

4.4.6 Fathom virtual 50 & 500 tosses of a coin simulation

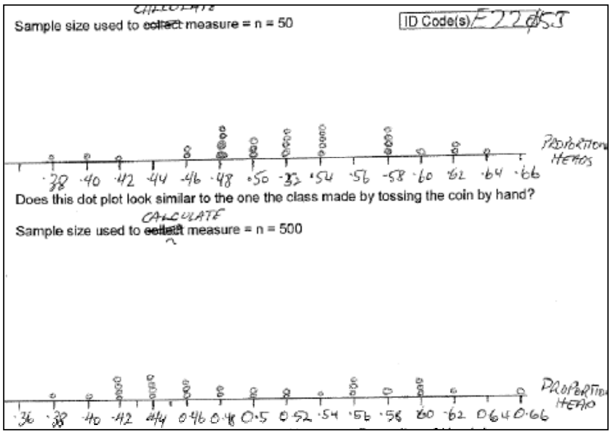
The objective of this activity was to build towards development of re-sampling techniques and of explicitly quantifying sample size by promoting two key concepts: (a) the spread of measures decreases as the sample size used to calculate the measure increases, and (b) the centre of the distribution of measures approaches the expected value as the sample size increases. The students assembled and used a Fathom coin simulation and were asked a series of questions that examined key features of the distribution. Methodology for the task is provided in Section 3.3.4.5, the worksheet is attached as Appendix A.12, and the results are presented in Table 4.20.

Table 4.20 shows that half of the boys and somewhat more than half of the girls provided multistructural responses, but that only one student provided a relational response that integrated all aspects of the task. This was an introductory task, and students were continuing to develop their understanding.

Table 4.20.
SOLO Evaluation of Students' Response First Fathom Coin Simulation

SOLO	Male		Female		
level	No.	%	No.	%	Criteria or exemplars
P	2	11 %	2	7 %	Irrelevant or no response, or not attempted
U	7	39 %	9	31 %	Simulation incomplete
M	9	50 %	17	59 %	All elements of activity completed. Four elements of task present
R	0	0 %	1	3 %	All elements presented in an integrated fashion. Correctly identify, provide or note – number of measures – expected proportion of heads – location of centre of data – narrower distribution at a large sample size – occurrence of distribution on a finer increment – centre of data likely to be closer to expected value at large sample size
Total	18	100 %	29	100 %	

Student E2205J provided a prestructural response to the task. The student's measures dot plot suggested that the student had generated data for a sample size of 50 tosses of a coin, but not for the larger sample size of 500; the student had imagined the distribution as being similar to the sample size of 50. The data for both dot plots were recorded at increments of 0.02, which is correct for a sample size of 50 because 0.02 corresponds to one head in a 50 tosses of a coin, but it is unlikely for a 500 tosses of a coin because the data should occur in increments of one tenth of that size at 0.002 (Figure 4.11).



CALCULATE
Sample size used to ~~collect~~ measure = n = 50

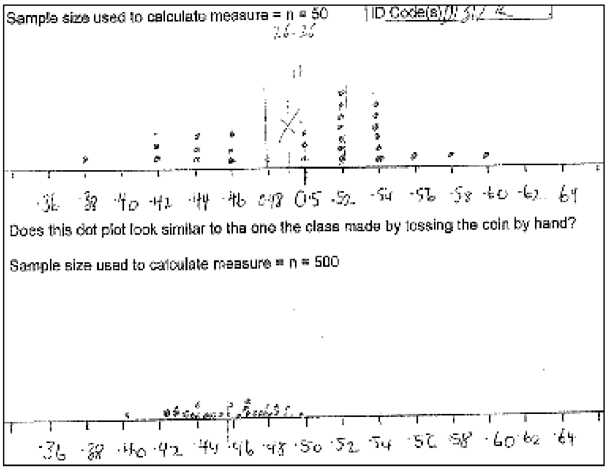
ID Code(s) E2205J

Does this dot plot look similar to the one the class made by tossing the coin by hand?

CALCULATE
Sample size used to ~~collect~~ measure = n = 500

Figure 4.11. Student E2205J, proportion of heads for samples size of 50 & 500 tosses of a coin, prestructural response.

Student D1312Z gave a unistructural response that demonstrated partial understanding of the key concepts. The student had constructed the simulation, but provided no robust evidence of collecting data at a sample size of 500 because the data centre was displaced to the left. The 500 tosses dot plot showed finer increments and a narrower distribution than the 50 toss of a coin. The student had identified correctly the centre within the misplaced data (Figure 4.12).



Sample size used to calculate measure = n = 50

ID Code(s) D1312Z

Does this dot plot look similar to the one the class made by tossing the coin by hand?

Sample size used to calculate measure = n = 500

Fig 4.12. Student D1312Z, proportion of heads sample size of 50 & 500 tosses of a coin, unistructural response.

Student H1112I provided a multistructural response where an appropriate response was provided for each item, but the responses were a series of disconnected answers that did not convey an integrated understanding. The student chose the mode as the average for both sample sizes. The centres of both distributions were located appropriately, the 50 tosses of a coin included one extreme data point and the student extended the scale, and the 500 tosses of a coin plot showed an appropriately narrow distribution (Figure 4.13).

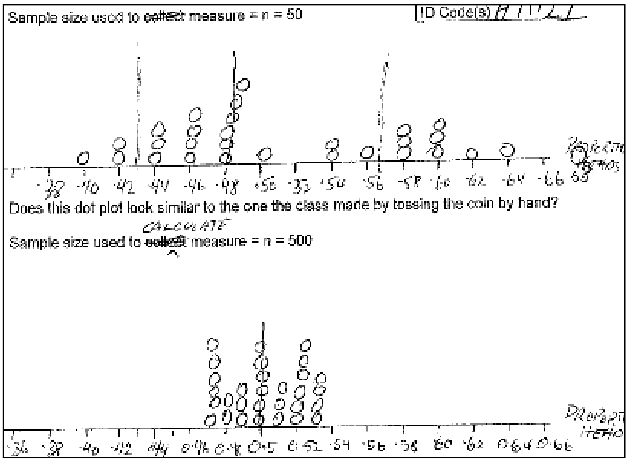


Figure 4.13. Student H1112I, proportion of heads sample size of 50 & 500 tosses of a coin, multistructural response.

Student G0709A provided a relational response (Figure 4.14). The student had chosen the mode in the sample size of 50, and had chosen either median or mean for the sample size of 500. The student correctly identified the number of measures collected, the effect of sample size on the centre and spread of the data, and the relationship between the expected value and the average. The responses conveyed a sense of integration of all aspects of the task. In response to whether the measures were likely to be closer or further away from the expected value the student answered correctly, and extended the discussion to the relationship between the sample size and the spread of the data.

> It will get closer. As you can see my second graph is a lot tighter than my first. [Student G0709A]

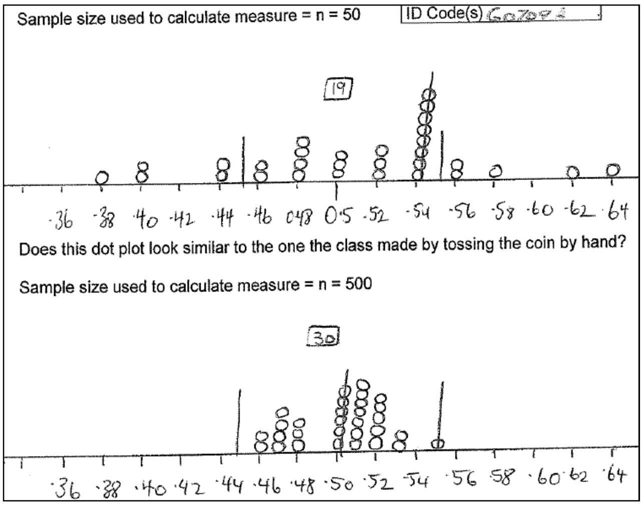


Figure 4.14. Student G0709A, proportion of heads sample size of 50 & 500 tosses of a coin, relational response.

4.4.7 Compare intuitive sense of 50 tosses of a coin with a Fathom coin toss

The objective was for students to compare their own intuitions of the distribution of a proportion of heads of a series of a 50 tosses of a coin in the Pre-test (Section 4.2.3) with the Fathom coin simulation. Methodology for the task is presented in Section 3.3.4.6, the worksheet is attached as Appendix A.12,Q.6 and Q.7, and the results are presented in Table 4.21.

Both male and female students over-estimated the number of proportions of heads that would occur within a given range; or, expressed alternatively, students predicted that the distribution of the proportion of heads was narrower (i.e., "less random") than occurs by chance. Table 4.21 shows that 28% of both male and female students predicted a larger number of proportions of heads would occur within the range than they actually observed in the Fathom simulation. Female students made a more accurate prediction than males with 34% of girls, but 17% of boys only, making a prediction similar to that observed with a Fathom simulation. Almost half of the male students (44%) did not provide a response: many boys did not give a response on the pre-test item and consequently these students were unable to make a comparison.

Table 4.21.
Comparison of Students' Prediction of a Coin Toss on Pre-test Item Q.4 with Their Data from a Fathom Simulation

Category of student response	Male students		Female students	
	No.	%	No.	%
Student predicted a larger number of proportions of heads within range than actually occurred / student over-estimated accuracy	5	28 %	8	28 %
Prediction close to actual distribution	3	17 %	10	34 %
Student predicted a lower proportion of heads within range than actually occurred / student under-estimated accuracy	2	11 %	4	14 %
No response or missing pre-test	8	44 %	7	24 %
Total	18	100 %	28	100 %

Students assessed their estimate as either accurate or inaccurate, rather than under or over-estimating the result. Many students, including approximately one third of female students, appeared to have intuitions that were supported by the Fathom simulation.

> My prediction was close, but not as close as if I'd said 28 out of 30 [Student S0403J]

> It was very close, but not quite right. [Student had predicted 20, but had observed 19]

Some students were satisfied with their predictions, but mathematically this confidence could not be justified. Their uncritical assessment would be unlikely to lead the students to question their own intuitions regarding this aspect of the random behaviour of a coin toss. Students G0709A, X0211G, and G2006J provide three exemplars. The students over-estimated the number of proportions of heads that would occur within a range.

> I was pretty close, although I still had a bit of a gap. [Student G0709A predicted 25 but observed 19]
>
> My prediction was pretty close. [Student X0211G predicted 25, but observed 21]
>
> Yes, it was sort of close. [Student G2006J had predicted 25, but observed 20]

Students who provided an inaccurate estimate tended to respond briefly without attaching any significance or analysing their response – the worksheet question alone did not provoke thoughtful analysis.

> No, it was bad. [Student S1510A predicted 25 within range, but observed 12]
>
> Not at all good. [Student N0909L had predicted 28, but observed 7]

4.4.8 Mt. Wellington cable car (naïve response)

The objective of the Mt. Wellington cable car task was for students to explore the sample size task for a public opinion survey either For or Against a well-known and controversial local development project. Students were placed in the role of responding to criticism of the sample size of 900 that was used. The task was presented to students first as a homework item, and presented again on the post-study assessment (Section 4.4.12) to allow an assessment of learning. Methodology for the task is presented in Section 3.3.4.7, the worksheet is attached as Appendix A.13, and the results for the item are presented in Table 4.22.

The proportion of students who provided a response was modest with 11 male students and 22 female students responding, and this may not be representative of the broader student cohort. The SOLO model was used for analysis, but the value of the item to this study lay principally in obtaining students' responses prior to formal study of the sample size model and contextual sampling tasks.

Table 4.22.
SOLO Evaluation of Students' Naïve Responses to Mt. Wellington Cable-car

SOLO level	Male		Female		
	No.	%	No.	%	Criteria or exemplars
P	2	18 %	4	18 %	Task not understood or irrelevant aspects considered.
U	6	54%	10	45 %	One or two elements of the criteria only, but not integrated.
M	3	27 %	6	27 %	Partial response, two criteria met.
R	0	0 %	2	37 %	Three criteria met and integrated.
EA	0	0 %	0	0 %	All four criteria met and presented in an integrated fashion – Calculates or recalls sample accuracy for the given sample size – Relates result accuracy to the alternative outcome – States outcome not certain – Representative and random sample (outside of task)
Total	11	100 %	22	100 %	

Two students provided a prestructural response stating simply that the sample size was reasonable or sufficient, but they did not provide a justification.

> The results are clear enough to state that a larger part of the community is against the cable-car. [Student H1112I]
>
> My survey shows are large enough amount of results. [Student S1510A]

Unistructural responses provided a single idea related to the question.

> […] should be a fair representation, not an exact opinion. [Student Y2907G]
>
> [….] even though the sample size was pretty small it gives a good general feeling. [Student N2103S]

Several students were clearly aware that a sample can only represent the population and recognised the role of chance and that additional sampling may change the result.

> […] there is a chance that it could improve [change] the result, but unless we survey all 200 thousand people there is still a chance [of the alternative outcome]. [Student M1306E]
>
> If the survey was to be done again over a larger number of people there [would] still be a chance that more will be against [Student R1207L]

Four students, all male, indicated that the sample was smaller than they would prefer instinctively – a result entirely consistent with the national and state polls presented in Section 4.2.4.

> That I agree [sample too small] and shall get more people to take the survey, 1100 to be exact, and add these results to the current results, but only if funded to do so. [Student N2701B]

One student, a male, provided the one example, with a justification, that the sample was larger than necessary.

> […] the sample size should be a bit smaller [in a population of one million] 900 would probably be just fine. [Student R1706D]

One student's multistructural response included the legitimate comment that a decision needed to be based on an overwhelming, rather than simple, majority. This lies outside the mathematical purpose of the task, and it suggests the student was thinking beyond the task and considering the implications of the result.

> I will say [….] 900 people is a smallish survey [….] would need a large popularity, at least 75% [in favour] to get the green light. [Student G0709A]

The belief in a sample as a proportion, in addition to being a representative part, of the population also suggests that a larger sample should be used for a larger population. The following student was reaching for a mathematical response.

> A sample of [900] in a city so small [as Hobart] is a respectable sample [….] your point [of a larger sample] would be understandable if we were surveying a larger city like Melbourne. [Student S0403J]

Several students attempted to use a mathematical approach in their analysis. If students explicitly considered sample size it was based upon a proportion of the population, or a consideration of the difference between the proportion For and Against.

> [….] 900 people is nearly 20 hundredths of Hobart's population. [Student Y2907G]
>
> I surveyed nearly 1 in 200 people (0.45%). [Student E3011L]
>
> [….] 9/2000 were surveyed [Student E1709S]
>
> The set of data is about 20% of the population and that is a larger percentage than normal [suggests student consider 10% an appropriate sample size] [Student D1312Z]

One student explicitly considered sample size and the difference between the proportion For and the proportion Against.

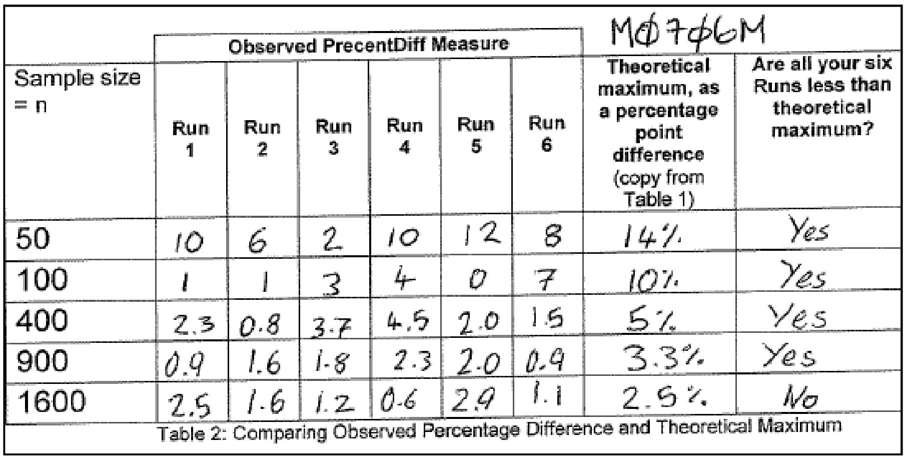
> Considering the very small margin, a larger sample size in another study may be warranted [….] to get a more precise result the sample size should be increased. [Student S1610J]

Student E1709S provided a sophisticated relational response in an approach more commonly used in sensitivity analysis. The student considered the consequences of additional sampling and the likelihood of changing the outcome of the survey. Engagement with the task was clearly high.

> Nine hundred is enough [….] and the opinion of the public [the result] isn't going to change by interviewing more people. 405 out of the 900 were For the project and 495 out of the 900 were Against. To change the result at least another [91] people would have to be interviewed and all of those 91 would have to be against [….] very unlikely. [Student E1709S]

4.4.9 Large population sample size model $e = \pm 1/\sqrt{n}$

The objective of this activity was to introduce the sample size model $e = \pm 1/\sqrt{n}$ as an estimator of the margin of error in random processes and examine the accuracy of the model using a process of statistical enquiry of a frequentist approach using Fathom. The sample size model $e = \pm 1/\sqrt{n}$ is as an estimator of the difference between the observed and expected values, otherwise known in the study as the margin of error. More formally, the model calculates a confidence interval within which 95% of measures of re-sampling will occur. Students calculated manually the margin of error, *e*, for the six sample sizes, *n*, of 49, 100, 400, 900, 1600 and 2500, and these margins of error were then compared with the data of the ObservedPercentDiff measure generated by a Fathom simulation. Students were provided with a functioning Fathom coin simulation, but modified the simulation using the formula and format editors. Methodology for this task is provided in Section 3.3.4.8, the task is attached in Appendix A.14, an example of the data generated by Fathom is presented in Figure 4.15, and the results are presented in Table 4.23.



MØ7Ø6M

	Observed PrecentDiff Measure							
Sample size = n	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Theoretical maximum, as a percentage point difference (copy from Table 1)	Are all your six Runs less than theoretical maximum?
50	10	6	2	10	12	8	14%	Yes
100	1	1	3	4	0	7	10%	Yes
400	2.3	0.8	3.7	4.5	2.0	1.5	5%	Yes
900	0.9	1.6	1.8	2.3	2.0	0.9	3.3%	Yes
1600	2.5	1.6	1.2	0.6	2.9	1.1	2.5%	No

Table 2: Comparing Observed Percentage Difference and Theoretical Maximum

Figure 4.15. Student M0709M's data for the large population sample size model.

The worksheet concluded with the question, “Overall do your think the $\boldsymbol{e} = \pm\mathbf{1}/\sqrt{\boldsymbol{n}}$ rule is a reasonable estimate of the Maximum Percentage Difference that is likely to be observed?” To complete the task successfully students based their assessment of the sample size model on the available evidence probabilistically. Students assessed the information in two sequential steps: Students first assessed the data their own simulation had generated, and second, through class discussion assessed all the data generated by the other members of the class.

Table 4.23.
SOLO Evaluation of Students’ Worksheet Large Population Sample Size Model

SOLO	Boys		Girls		
level	No.	%	No.	%	Criteria or exemplars
P	5	24 %	5	15 %	No explanation or not completed.
U	8	38 %	6	18 %	Simple agreement or disagreement supported by single justification. Evidence simulation constructed correctly.
M	6	29 %	11	32 %	Interpretation consistent with available evidence, but not all the available evidence is used. Evidence of simulation completed.
R	2	9 %	12	35 %	Comprehensive interpretation based on and consistent with all the available evidence.
Total	21	100 %	34	100 %	

Student E3011L provided a unistructural response. The student’s own simulation data yielded two results higher than that predicted by the rule, but the student nevertheless concluded the rule as a “reasonable estimate,” but not a rigidly enforced or precise “rule.” The student recognised the limitations of the sample size model.

> I think that it is a reasonable estimate but it is not to be taken as a rule or a precise marking. [Student E3011L]

Student M1306E provided a multistructural response where the data at the different sample sizes were compared with the model, and the student agreed with the question that the model was a reasonable estimate. The student was aware of chance, and of making a probabilistic decision. The rule was then applied to consider whether the sample was fair. This was potentially a misapplication, but it did indicate thinking beyond the immediate task and an attempt to link the activities conducted earlier in the study of the fairness of the die. On balance this was a multistructural response.

> Yes, I do. Most of the results are under the $e = \pm 1/\sqrt{n}$ rule, and there is a chance factor. The results that were over could be because of chance, so I do think it is a fair sample. [Student M1306E]

Student E2909G provided a second example of multistructural response. The student critically reflected on the data generated by the simulation. The student examined the data at different sample sizes and thought the rule was less accurate at small sample sizes was apparently comfortable with the model's limitations. The more accurate expression "percentage points" is used rather than percentage.

> Yes [the rule is accurate] because most of the sample sizes – theoretical maximum was accurate, and where it wasn't accurate it was only out by a few percentage points. The most extreme case was only out by two percentage points in 50 runs [….] was expected with the smaller sample size anyway. [Student E2909G]

Student S0403J provided a relational response where the student considered both her own data and the data generated by the class. The task asked only whether any simulation results exceeded the (maximum) margin of error calculated by the sample model, but the student extended the response and considered both the number and proportion of the individual results that exceed the maximum – an example of formal analysis and mathematising that was a key objective of the study.

> Yes, there were few exceptions or none at all according to my results. When compared with the class only two or three results don't match. 3 out of 5 sample sizes had runs less than the theoretical maximum, which is 60% (although they are my results only). [Student S0403J]

Student S0412N had responded to the task, but continued to demonstrate considerable reservation about probabilistic decision-making and the study of statistics. The student, as part of a multistructural response conceded that the model may be valid sought to apply mathematical rigour to sample size.

> It does [the rule is valid] in a sense, [but] the whole statistical thing is very confusing because I like things that are accurate and dealing with maths that isn't truly accurate is unnerving. [Student S0412N]

4.4.10 Federal election survey: Howard and Rudd (Post-study test Q. 2)

The objective of this task was to provide students with an opportunity to apply their understanding of sample size to a media report of an opinion survey conducted prior to the 2007 Australian national election. The title of the item was derived from the leaders of the two major Australian federal political parties at the time of the study.

Methodology for the task is provided in Section 3.3.5.2, the worksheet is attached as Appendix A.15, Q.2, and the results are presented in Table 4.24.

Table 4.24.
SOLO Evaluation of Students' Post-study Test Item Q.2 Howard and Rudd Election Survey

SOLO	Boys		Girls		
level	No.	%	No.	%	Criteria or exemplars
NR	1	5.5 %	1	2.9 %	No response, not attempted.
P	5	27.8 %	6	17.1 %	Irrelevant response.
U	5	27.8 %	10	28.6 %	One or two element of the criteria only, but not integrated.
M	3	16.7 %	11	31.4 %	Partial response, at least two criteria met.
R	4	22.2 %	7	20.0 %	All three criteria met and presented integrated. – Recalls accuracy for a given sample size, or calculates the margin of error using the sample size model – Relates result accuracy to the alternative outcome – States outcome not certain
Total	18	100.0 %	35	100.0 %	

Student S0412N provided a prestructural response that included irrelevant material "people's opinions change" and that "some are not truthful" – certainly valid comments, but arguably this also represents material that confounds and distracts.

> No, people's opinions change and some are not truthful. [Student S0412N]

Student E2909G provided a unistructural response that at least recognised that a survey result is subject to natural variation, but incorrectly stated that the sample size is not large enough.

> No, because the survey has natural variation that could change the results and the sample size isn't big enough for a conclusive result. [Student E2909G]

Student E1709S provided a strong, multistructural response, but one that did not explicitly demonstrate use of the sample size model. The accuracy of the survey is related appropriately to the likelihood of the survey reporting a different outcome.

> Most likely yes, because the percentage [results] will range 44.5–49.5% and 50.5 – 55.5%, which makes the chance of the highest percentage [the outcome] changing unlikely, but a small possibility if there was another survey because they are so close. [Student E1709S]

Student S0403J provided a relational response that included calculating a survey's accuracy of +/- 2.5% at a sample size. The student used the qualifying expression "who is most likely to win."

> Using the $e = \pm 1/\sqrt{n}$ rule I calculated the results could vary +/- 2.5%. If you [….] subtracted 2.5% Kevin Rudd would still be favoured by a majority. I don't think a survey could conclusively show who wins the election, but I think it shows who is most likely to win. [Student S0403J]

4.4.11 Mathematics of the sample size model (Post-study assessment Q. 3 b & c)

The objective of these two items was to assess students' understanding of the mathematics within the sample size function. The first item, Q.3 (b), asked students to calculate the margin of error reworded as the "range" for a sample size of 200. The second item, Q.3 (c), reversed the conventional order of calculation and asked students to determine the sample size for a given range of measures read from a dot plot (Figure 4.16). Methodology of the task is presented in Section 3.3.5.4, the item is attached as Appendix A.15, Q. 3 b & c, and the results are presented in Tables 4.24 and 4.25.

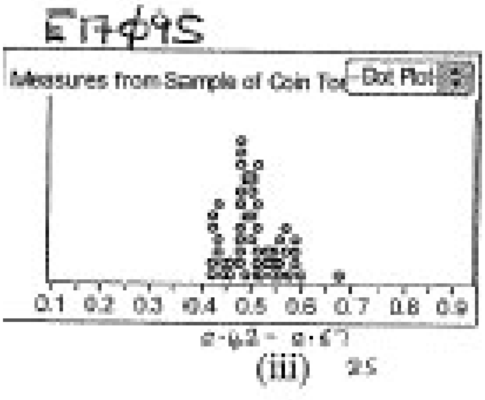


Figure 4.16. Using the distribution of heads to estimate the sample size, Q.3 (c).

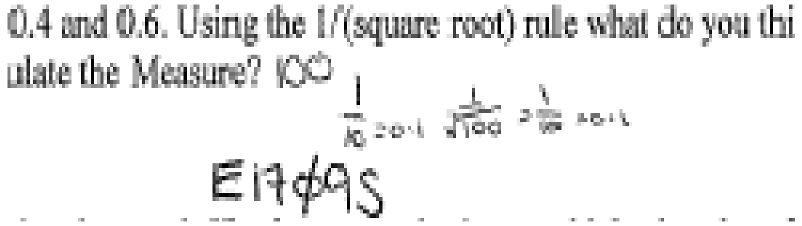
Table 4.25 shows that a larger proportion of boys than the girls were procedurally correct on Q 3.b. A procedurally correct response substituted a value of 200 into the sample size model and calculated correctly the range of (0.43, 0.57), a partially correct response was demonstrated by an attempt to substitute the value and partially complete the calculation, and an incorrect response was demonstrated by a student unable to use the model procedurally. Two thirds of the girls (66%) and 39% of the boys gave either an incorrect response or no response.

Table 4.25.

Students' Reponses to Post-study Test Item Q. 3(b) Use of the Large Population Sample Size Model to Determine the Spread of the Measures

Students' responses	Male students		Female students	
	No.	%	No.	%
Fully correct	6	33 %	9	26 %
Partially correct	5	28 %	3	9 %
Incorrect	4	22 %	8	23 %
No response	3	17 %	15	43 %
Total	18	100 %	35	100 %

Item Q. 3 (c) required, if solved formally, algebraic manipulation of a function more complex than normally considered at Year 9. The item can be solved in a number of ways. To solve the item formally students noted that the question stated that most of the measures lay between 0.4 and 0.6 (Figure 4.16), recognised that this range was equivalent to 0.5 ± 0.1 and a margin or error e = 0.1, solved the sample size model for $n = (1/e)^2$, and substituted e = 0.1 into the function. Other strategies were to recall from the classroom discussion that a sample size of 100 has a margin of error of ± 0.1 or to have the number-sense to recognise the relationship within the sample size function between the sample size of 100 and the margin of error of 0.1.



0.4 and 0.6. Using the 1/(square root) rule what do you thi

ulate the Measure? 100

$\frac{1}{10} = 0.1 \quad \frac{1}{\sqrt{100}} = \frac{1}{10} = 0.1$

E17Ø9S

Figure 4.17. Student E1709S, calculation of sample size based on the data spread.

None of the students who gave a correct response manipulated the sample size function algebraically, but they used their number-sense and their knowledge of sample size from the class to determine the correct sample size. Table 4.26 shows that 28% of male and 17% of female students provided a correct response. A large proportion of the female students (60%) did not provide a response, and these students may have been discouraged from responding because of the warning that the question was "tough."

Table 4.26.

Students' Responses to Post-study assessment Q. 3(c) Reversing the Conventional Order of Calculation and Determining the Sample Size from the Spread of the Measures

Students' responses	Male students		Female students	
	No.	%	No.	%
Correct	5	28 %	6	17 %
Incorrect	9	50 %	8	23 %
No response	4	22 %	21	60 %
Total	18	100 %	35	100 %

The students found the application of the sample size function to tasks beyond those encountered directly in class challenging. One of the objectives of the study was to mathematise sample size, but by the conclusion of the classroom study students' knowledge of the mathematics of the model was limited.

4.4.12 Mt. Wellington cable-car (Post-study test Q. 5)

The objective of the Mt. Wellington cable-car activity was to provide students with an opportunity to demonstrate the skills and knowledge of sample size in a contextual task. The task was offered to students first as a homework item (Section 4.4.8) prior to the introduction of the sample size model to assess their naïve response and here as a post-study item. Methodology for the task is presented in Section 3.3.5.6, the task is presented in Appendix A.15, Q. 5, and the results are presented in Table 4.27.Students' development from the naïve to the post-study assessment is considered subsequently in Section 4.4.14.

To complete the task successfully students performed a series of two steps to choose a sample size: consider the interrelationship between sample size and the accuracy of the survey; and integrate the sample size, the accuracy of the survey, and the significance of the result relative to an alternative survey outcome. The task as presented offered scope to demonstrate higher order thinking.

Table 4.27 shows that 56% (10/18) of male and 74% (26/36) of female students provided a multistructural response or above. A significant difference occurred between the proportion of male and female student students who gave a high level of response, with 48% (17/36) of the females but only one male student (6%), providing a relational or higher level of response. This suggests that many of the female, but only one of the male students had developed sufficient understanding of the concepts to apply the sample size model to a contextual task. The four criteria used for assessment are presented in full for the extended abstract response.

Table 4.27.
SOLO Evaluation of Students' Post-study Assessment Mt. Wellington cable car

SOLO	Male		Female		
level	No.	%	No.	%	Criteria or exemplars
P	4	22 %	7	19 %	Irrelevant, no response, or not attempted.
U	4	22 %	3	7 %	One element of the criteria only.
M	9	50 %	9	26 %	Partial response, two criteria met.
R	1	6 %	13	37 %	Three criteria met.
EA	0	0 %	4	11 %	All four elements of the criteria are met. − Calculate or recall sample accuracy for the given sample size − Relates result accuracy to the alternative outcome − States outcome not certain − Representative and random sample (outside of task)
Total	18	100 %	36	100 %	

Student N0106D provided a unistructural response that demonstrated awareness that a representative sample should be used. Justifying a sample size, to quote the student, as a "fair amount" had no mathematical basis. The student made the legitimate point that not all surveyed may have had an opinion, but such a comment in a low level response was irrelevant to the mathematical concepts being studied.

> Not everyone living in Hobart would have an opinion. You asked a range of different people. 900 is a fair amount to be surveyed. [Student N0106D]

Student Y0706J provided a multistructural response. The student demonstrated correct calculation of the sample size rule and indicated an awareness of the significance of the survey result, but did not explicitly relate survey accuracy to the survey result.

> If we [used] $1/\sqrt{n}$ ± 3.3%, it will still show that more people are Against. We don't need to increase the sample size. [Student Y0706J]

Student E3011L provided a relational response that included a calculation of the accuracy of the survey for the given sample size and related the accuracy to the alternative outcome. The terminology was used loosely. The response was particularly significant because the student also calculated the sample size as a fraction of the population. The student used the calculation 9/2000 in precisely the same manner used in the student's naïve response in the homework item. This suggests the persistence of the sample as a proportion of the population, and that the student was not fully convinced of the value of the sample size model.

> $1/\sqrt{n}$ = 1/ √900 = 3.3%. The results are accurate to about 3% therefore the range of people 'For' the cable car are between 42–48%, so therefore the majority of people are against it. I surveyed 9/2000 people from all walks of life (0.75%) [Student E3011L]

Student M1306E provided an extended abstract response. The student noted the use of a representative sample, acknowledged that a survey is inherently less accurate than a census, but that a survey allows resources of time and cost, that a sample size of 900 allows an accuracy of 3.3%, and that the result with this accuracy will not change the outcome.

> [….] thorough survey people of all ages, races, and areas were tested. The only way to get an exact answer is to survey [everyone, and] we simply have neither the time nor funds. A sample size like we tested there is a natural variation of about ± 3.3% band. […] When looking at 45% for the band is (45-3.3%) = 41.7% to (45 + 3.3%) = 48.3% […] If we tested a different 900 people the result could change but there can't be a majority 'For' the cable-car. [Student M1306E]

4.4.13 Sample size for a National and state election survey (Follow-up test)

The National and state election survey item, first presented to students as a homework item (Section 4.4.8) to give their naïve response, was re-presented to the students approximately two months after the conclusion of the classroom study to determine whether any development of understanding of sample size that had occurred in the classroom was sustained. Students did not receive any tuition between the conclusion of the classroom study and the follow-up test. Methodology for the task is presented in Section 3.3.5.8, and the item is attached as Appendix A.17. The results and the criteria used to assess students' responses are presented in Table 4.28.

Table 4.28.
SOLO Evaluation of Students' Post-study Responses to National and State Election Survey Item

SOLO	Males		Females		
level	No.	%	No.	%	Criteria or exemplars
P	0	0 %	0	0 %	Task not understood or irrelevant aspects considered.
U	7	39%	5	17 %	One or two elements of the criteria met only (see below), but elements not integrated.
U_TM	6	33 %	7	23 %	
M	1	5 %	12	40 %	Partial response, two criteria met(see below).
M_TR	4	22 %	6	20 %	
R	0	0 %	0	0 %	Three criteria met and integrated.
EA	0	0 %	0	0 %	Comprehensively integrates all criteria. – formally calculates error using $e = \pm 1/\sqrt{n}$ – relates sample size to accuracy – aware that sample size is independent of population size – random and representative sample – sample as an imperfect but accurate representation of a population – rejects 10% of population as unfeasible sample size – calculates or recalls sample accuracy for the given sample size – relates result accuracy to the alternative outcome – states outcome not certain – notes importance of representative and random sample (outside of task)
Total	18	100 %	30	100 %	

Student S1808B provided a unistructural response. The student nominated the preferred sample size of 1,500 but did not provide an explanation or justification. The student's understanding could be explained by one concept only: use a sample size of 1500 when sampling very large populations.

Student R2408I also provided a unistructural response. The student used an inconsistent strategy using 10% of the population for the Australian survey and a sample size of 15,000 for the Queensland survey – the student was unable to relate the 10% with an actual sample size. The student sought to maximise accuracy rather than consider meaningful accuracy. The sample size for Queensland was chosen as "reasonable," but the student failed to provide an explanation.

> …the sample size needed to be as large as possible to achieve more accurate results ….In relation to the population of Queensland 15,000 seems like a reasonable number. [Student R2408I]

Student E2611G provided a multistructural response. The student nominated a sample size of 15,000 for both surveys, but considered accuracy informally only. A sample of 10% was considered more accurate but not feasible. The student noted correctly that at a large population the sample size is independent of the population size. The student stated incorrectly that the state election would be more accurate, which suggests that the student continued to consider a sample as a proportion of the population.

> If they surveyed 10% of the population the percentage might be more perfect but it would cost too much, 1,500 and 150 people are too little and the range would not be accurate. Just because the population is smaller the sample size doesn't have to be smaller. The survey would be more accurate than the federal survey. [Student E2611G]

Student N2306C provided a transitional M_TR response. The student nominated the same sample size for both surveys and recalled the expression "margin of error" used in the classroom study. The preferred sample size of 1500 was used, which the student could have recalled from memory, but this also demonstrates confidence in using a small sample size.

> I thought that 1,500 would be enough to cover a whole range of opinions. The amount of people would also be enough to have a reasonably small margin of error. [Student N2306C]

Student M0706M, as part of a transitional U_TM response, seemed troubled by the model that calculated sample sizes that may be larger than the population sizes they encountered in earlier mathematics courses. The study did not consider small population sample size because of the complex mathematics involved. The large population sample model has application only for large or infinite populations, whereas the 10% rule may seem to have application in all small and large population surveys.

> Surveying 10% of the population also works well no matter what the size of the population is. [Student M0706M]

4.4.14 Students development on two contextual tasks (a) Mt. Wellington cable-car survey and (b) national and state election survey

This sub-section considers students' longitudinal development of understanding of sample size in context of public opinion surveys where students' naïve understanding prior to tuition is compared with their understanding on the final assessment. Two items are examined here: the first is the Mt. Wellington cable-car, and the second is the national and state election. The Mt. Wellington cable-car item was presented initially to students immediately prior to the introduction of the sample size model and re-presented as an item on the final assessment at the conclusion of the teaching unit. This item assessed students' development of understanding as would occur in a classroom teaching unit where assessment was conducted immediately at the conclusion of the teaching unit. The national and state election items were presented to students first as part of the pre-test and two months after the conclusion of the classroom study; these were designed to assess students' long-term development of understanding that students would use beyond the classroom.

On the first item, the Mt. Wellington cable-car, which was assessed at the conclusion of the classroom unit of work, the level of response of both male and female students increased substantially. The females demonstrated a substantially higher performance than the male students (Table 4.29). Male students predominantly gave a unistructural response (54%, 6/11) prior to the study, but on the post-study assessment the proportion who gave this response decreased to 22% (4/18), and the proportion of males who gave a multistructural response increased from 27% (3/11) to 50% (9/18). Prior to the study female students predominantly also gave a unistructural response (45%, 10/22), but on the post-study assessment the proportion of females who gave this response decreased to 7% (3/36) and the proportion who gave a relational response increased to 37% (13/36). The proportion of students who gave a low, prestructural, response – approximately 1/5th of both male and female students – was unchanged across the classroom study. This suggests that a substantial proportion of students showed development of understanding of sample size of the Mt. Wellington cable-car, but that 1/5th of students displayed no development.

Table 4.29.
Comparison of Students' Naïve and Post-study Responses Mt. Wellington Cable-car

SOLO level	Male students		Female students	
	% Prior (n=11)	% Post-study (n=18)	% Prior (n=22)	% Post-study (n= 36)
Prestructural (P)	18 %	22 %	18 %	19 %
Unistructural (U)	54 %	22 %	45 %	7 %
Multistructural response M)	27 %	50 %	27 %	26 %
Relational (R)	0 %	6 %	0 %	37 %
Extended abstract (EA)	0 %	0 %	0 %	11 %
Total	100 %	100 %	100 %	100 %

In the second item, the national and state election survey, Table 4.30 presents students' long-term development of understanding of sample size that students would take outside the classroom. SOLO transitional responses were included in the analysis because students' responses lay within a narrow range, and the transitional responses allowed subtle distinctions to be made. Students showed modest development only. No student provided above a transitional multistructural relational M_TR response. The $e = \pm 1/\sqrt{n}$ rule was recalled by two students only, so the study's objective of providing a convenient and mathematically correct alternative to the 10% of the population rule was judged unsuccessful. The most significant development (refer Section 4.4.13) was the informal consideration of accuracy by female students and a greater appreciation of sampling issues.

Table 4.30.
SOLO Analysis of Students' Responses National and State Election Survey: A Comparison of Student Pre-Test and Follow-up Test Responses

SOLO level	Male students		Female students	
	% Pre-test (n=21)	% Follow-up (n=18)	% Pre-test (n=35)	% Follow-up (n=30)
Prestructural	0 %	0 %	0 %	0 %
Unistructural (U)	48 %	39 %	41 %	17 %
U_TM	24 %	5 %	9 %	10 %
Multistructural (M)	24 %	33 %	34 %	41 %
M_TR	0 %	22 %	0 %	21 %
Relational	0 %	0 %	0 %	0 %
Extended abstract	0 %	0 %	0 %	0 %
Incomplete response	4 %	0 %	16 %	0 %
Total	100 %	100 %	100 %	100 %

Table 4.31 examines students' choice of sample size. Both male and female students gave a low-level response to the item. The proportion of male students who gave the preferred response of a sample size of 1,500 increased from 9% to 17%, and the proportion of female students who gave this response decreased from 21% to 3%. Two students only made a reference to the sample size model, but neither of the two attempted to apply it. The most significant change was the substantial decrease in the proportion of students who preferred a sample "10% of the population" and the substantial increase in the proportion of students who preferred a sample size of 15,000. This suggests that students may have simply replaced a sample size 10% of the population with the next largest sample size alternative offered. A sample size of 10%

of the population is not practicable, and students may have calculated the actual sample size 10% of the population would represent. Nevertheless, approximately one third of both male and female students continued to prefer a sample size of 10% of the population from amongst the alternatives offered.

Table 4.31.
National and State Election Survey Sample Size: A Comparison of Students' Pre-Test Responses with the Follow-Up Test Responses

Sample size strategy	Male students		Female students	
	% Pre-test (n=21)	% Follow-up (n=18)	% Pre-test (n=35)	% Follow-up (n=30)
(a) About 10% of the population	62 %	34 %	55 %	30 %
(b) 15,000	29 %	50 %	18 %	67 %
(c) 1,500*	9 %	17 %	21%	3 %
(d) 150	0 %	0 %	0 %	0 %
No response	0 %	0 %	6 %	0 %
Total	100 %	100 %	100 %	100 %

* Note: The accepted sample size when sampling from large populations

Table 4.32 presents the data for whether students adopted a consistent or inconsistent sample size strategy on the pre- and follow-up tests. A consistent strategy was defined in Section 3.3.2.4 as the same strategy for both the national and the smaller population state opinion survey. The proportion of male students who adopted an inconsistent strategy (largely a 10% of the population for the national survey and a numeric strategy for the smaller state population) decreased marginally, and the proportion of female students who adopted this strategy on the follow-up test was not significantly different from the pre-test. The proportion of both male and female students who would use a consistent 10% strategy decreased substantially. The male students favoured a consistent and constant numeric sample size strategy, and the female students adopted a numeric strategy, but would use a smaller sample size for a smaller population. The proportion of female students who gave an incomplete response decreased and all female students who participated in the follow-up test gave a complete response, which suggests that both males and females were now at least sufficiently confident to give a response.

Table 4.32.

An Examination of Whether Students Used Consistent or Inconsistent Strategies on the Pre-test and the Follow-up Test for National and State Election Opinion Survey Sample Size

Sample size strategy	Male students		Female students	
	% Pre-test (n=21)	% Follow-up (n=18)	% Pre-test (n=35)	% Follow-up (n=30)
Inconsistent strategy, combination of "10% of population" and numeric.	28.6 %	22.2 %	15.1 %	13.3 %
Consistent "10% of the population" for both national and state polls	28.6 %	16.7 %	45.5 %	16.7 %
Consistent numeric sample size with a smaller sample for the smaller state population	33.2 %	27.8 %	18.2 %	60.0 %
Consistent numeric strategy of 15,000 for national and state election	0.0 %	16.7 %	0.0 %	6.7 %
Consistent numeric sample size of 1500 for national and state election*	4.8 %	16.7 %	3.0 %	3.3 %
Incomplete or no response	4.8 %	0.0 %	18.2 %	0.0 %
Total	100.0 %	100.0 %	100.0 %	100.0 %

* The preferred strategy

4.4.15 Students' post-study questionnaire items

Students' conceptual understanding of sampling, sample size, and error was explored through a series of nine items: Items 10, 11, 13, 17, 19, 20, 23, 26, and 28. None of the items involved calculation or context, and many items were not explored explicitly in the classroom study. Students were assessed as either correct or incorrect. Methodology for the questionnaire is presented as Section 3.4, the questionnaire is attached as Appendix A.18, and the results are presented in Table 4.33.

Items 28, 17, 19, 26, and 13 examined students' understanding of sample size. The data for these items are purposefully presented in what was thought by the researcher as the order of increasing complexity and sophistication of the concepts, rather than the same order given to students in the questionnaire. Presenting students' responses in order of item complexity is designed to place students' development on a spectrum from naïve to sophisticated understanding. Female students' performance on these items tasks followed task complexity inversely, i.e., as the task complexity increased students' performance decreased. The male students' performance also followed task complexity, with the exception of Item 26 where their understanding of the significance of the margin of error (expressed as band of variation) in interpreting the result of a survey was weaker than their understanding of the other items.

Item 28 “If the survey is likely to be close I may need to take a larger sample size” and Item 17 “When conducting a survey the accuracy I need determines the sample size I must use” were two questions that could be answered correctly from general knowledge, without referring specifically to the material presented in the study. Students’ response were correspondingly high on Item 28 with 100% of male and 91.4% of female students correct, but on item 17 only 66.7% of male students and 82.9% of female students gave the correct response.

The three items 19, 26, and 13 required students to integrate sample size, survey accuracy, and the sample size model. Item 19 “If the results of a representative survey are likely to be clear-cut (e.g., 90% YES & 10% NO) then the sample size could be reduced,” considered a highly polarised community attitude survey that was not considered in the study, but nevertheless it demonstrated that less accuracy is required when the outcome is clear; approximately three quarters of the male (72.2%) and female (74.3%) provided the correct response. The male students’ performance on Item 26 that considered the consequences of the survey accuracy was low, but this topic was not included in class discussion. Item 13 was designed to assess students’ application of the algebraic sample size formula by interpreting the mathematical model verbally.

Items 20, 10, 23, and 11 explored students’ fundamental beliefs regarding sampling. In Item 20 half of the male students preferred the unambiguous nature of a census, describing a census as “useful.” Item 10 showed a persistent preference for a sample size of 10% of the population amongst one third of both male and female students. The erroneous belief that a sample size must be related to the population size continued to be widely held by students: Item 23 shows that only 11.1% of the boys and 8.1% of the girls responded correctly that a national opinion poll did not require a larger sample size than state opinion poll. Item 11 considered the cost/benefit of increasing the sample size, but expressed in the “measures” terminology of the Fathom simulation; 50% of boys and 27% of the girls responded correctly.

Table 4.33.

Evaluation of Students' Post-study Questionnaire Sample Size Concepts Items

Post-study questionnaire item	Male students Correct response (n=18)		Female students Correct response (n=37)	
	No.	%	No.	%
28. If the survey is likely to be close I may need to take a larger sample [Correct response: agree / strongly agree].	18	100 %	32	91.4 %
17. When conducting a survey the accuracy I need determines the sample size I must use [agree / strongly agree].	12	66.7 %	29	82.9 %
19. If the results of a representative survey are likely to be clear-cut (e.g. 90% YES & 10% NO) then the sample size could be reduced [agree / strongly agree].	13	72.2 %	26	72.9 %
26. A maximum band of variation of +/- 4% is not important when the survey showed 45% FOR and 55% AGAINST [disagree / strongly disagree].	3	16.7 %	23	62.3 %
13. Doubling the sample size halves the error [disagree / strongly disagree].	12	66.7 %	13	37.8 %
20. Only one survey is useful, and that is when everyone is asked (a census) [disagree / strongly disagree].	9	50.0 %	23	70.3 %
10. When I am conducting a survey I will almost always use a sample size of 10% of the population [disagree / strongly disagree].	6	33.3 %	13	37.8 %
23. Surveys prior to national elections are more expensive to conduct than surveys prior to state elections because many more people must be sampled [agree / strongly disagree]	2	11.1 %	3	8.1 %
11. It is always worthwhile to increase the sample size used to calculate a measure [disagree / strongly disagree]	9	50.0 %	9	27.0 %

4.4.16 Detailed study cumulative proportion of heads

Part F of the detailed study explored sample size and measurement error, but it varied the classroom learning sequence somewhat and examined students' responses to a cumulative proportion of heads of a coin toss. The three elements of situated abstraction of construction, recognition, and building-with provided the framework that supported analysis of students' responses. Methodology for the task is presented in Sections 3.7.7 ,and the worksheet is attached as Appendix D.1 Part F.

Part F of the detailed study examined students' interpretation of a graphical representation of cumulative proportion of a coin toss. Students were provided with a Fathom coin simulation of a sample collection containing four attributes: the number of times the coin was tossed, the outcome of the coin tosses of either a "H" or "T", a running tally of heads, and the cumulative proportion of heads.

The task began by first establishing that students understood the data set by being able to interpret correctly the attributes – this was the first, and seemingly essential, element of construction. Students identified accurately the simulation attributes; the one exception was student S1001J who read the attribute names in a mechanical fashion and included the underscore character that Fathom required to produce a one word attribute title. The student interpreted "no." as the word "not," rather than short for "number." The attributes were not understood immediately by this student.

> Toss, times underscore toss, no underscore heads, prop underscore heads [….] how many times it has not appeared heads. [Student S1001J]

The elements of construction and recognition could run together. When students interpreted the graph their attention was initially attracted by the erratic behaviour of the graphed data, and these initial observations were not necessarily explicitly mathematical.

> The graph it is all over the place and then levels out. [Student Y1504L]
>
> Starts up the top [proportion of heads is one], and sort of goes down below the line…it like…it goes up and below the line and it keeps going […] it's always staying near the line. [Student T0612M]

Students' thinking became progressively more sophisticated and mathematical as other features, including scale, were incorporated into the analysis. Repeatedly running the simulation provided the opportunity for students to build-with their existing knowledge to develop and subsequently test proto-theories, such as categorisation of the simulation output or some feature that occurred. This was a part of sense-making that provided an essential foundation without which a more formal mathematical approach may have had little meaning. One such proto-theory was that the initial value of the proportion of heads was either one or zero, which the students expressed informally as top or bottom. A second proto-theory was that if the observed proportion of heads started above the expected proportion of heads of 0.5 it tended to stay above 0.5; a proto-theory that was not mathematically correct and one which the student subsequently discarded.

> Like it always starts up the top or down the bottom and it edges towards the middle as you get closer to the end [Student T0612M]
>
> When…it has a higher proportion of heads it has a higher proportion of heads for the entire 600 counts and when it is lower it is the same for all 600 counts so it doesn't go above or below the line usually either all on top of the line or all below. [Student E1709S]

One student's general observation of the cumulative proportion of heads became more formal as the student sought to categorise the behaviour of the plots.

> In sort of like the 100 to 250 [sample size] it seems to go up and down on all the graphs, so it doesn't sort of stay straight there, and it straightens out after the 300 mark or something. It only does three things [categorisation] it will either go just below or just above or below the line [the 0.5] or spot on sort of thing… three more samples that it's a bit bigger than usual…just before it went under, then just middle, then a bit above over, now it has gone a bit further than that. [Student S1001J]

The analysis became more formally mathematical as the researcher asked the students to compare the observed value at two different sample sizes (100 and 600 were chosen), read the values and calculate the difference between the observed and expected at different sample sizes.

> With mine it is really …it is a lot larger [at 100] than it is at 600…it just seems the bigger the sample size the closer it gets to the line [expected] and it doesn't keep jumping about as it does. [Student Y0304T]
>
> At 100 it is about the same as it is at 600, but overall [when re-sampled a few times] it gets smaller …the difference gets smaller. [Student Y0706J]

None of the students quantified the difference without prompting. Students were now building-with their existing knowledge of the graphs' behaviour and their mental arithmetical skills. The formal calculation was a challenging task that involved decimal fractions smaller than one. Using cues from the other student one student addressed the task by first comparing the observed values at different sample sizes, and then calculating the difference between the observed proportions of heads and the expected value. This was a challenging mental arithmetical task. Student T0612M demonstrated both greater understanding of the task and competence in the calculation.

> S1001J – It is zero point 55 at 100 it is 0.59 and at 600 it is 0.52, so it changes a bit, but it isn't a huge change [doesn't appreciate significance of the number]…it gets smaller.
>
> Researcher – So is the difference getting smaller? Generally…I know you had a few results…
>
> S1001J – It gets smaller.

> Researcher: If you had to put a number on it …the size of that gap …and you had to put a number on it…
>
> S1001J: – You mean from 0.5?
>
> Researcher – Yes, the difference between the observed value – the "squiqqly" line and the 0.5 the expected value…
>
> S1001J – It's zero point 25 [is student still subtracting the observed values at 100 and 600?]
>
> T0612M – Zero point one two [corrects himself] zero point zero one two [correct 0.012]
>
> S1001J – Zero point zero two five [correct 0.025]

The preceding extracts examined students' thinking about one aspect of re-sampling: their observations and interpretation of repeated runs of the coin-toss simulation that sought to demonstrate the relationship between the cumulative proportion of heads and sample size. Students' initial informal observations became progressively more sophisticated and mathematical as the features of the graphs were incorporated into their analysis as students constructed and assembled the information. Students' proposed proto-theories as they sought to abstract meaning from the tasks. The researcher guided students towards a formal measurement of the relationship between sample size and the proportion of heads, but students found the mental arithmetic involved challenging. Students first interpreted the underlying data set that included identifying the attributes and developed a global perspective of the data. Initial observations of the cumulative proportions of heads graph included the informal observations of the data's behaviour (e.g., "the trend straightens out after the 300 mark" [Student S1001J, Appendix D.4]), or informal proto-theories of the behaviour of the trend graph (e.g., "when it has a higher proportion of heads, it has that for the entire 600 tosses" [Student E1709S], Appendix D.5). Proto-theories, which are elements of schemes, were exposed using Fathom simulation – the theories could also be tested using Fathom and discarded if false. The detailed study interview transcripts showed five instances where students only stepped beyond the informal analysis to a more formal mathematical approach when prompted and questioned by the researcher,e.g., "plus or minus speak? […] If you had to put a number on it?" [Researcher, Appendix D.2]. Without guidance from the researcher the students tended to meander and focus on unimportant or irrelevant features such as being pre-occupied solely on the erratic nature of the cumulative proportion of heads.

Students struggled to calculate mentally the difference between the observed and expected values: hardly surprising given the fractional decimals involved. The calculation was also at two different samples sizes – two slices – of the plots, and students needed to ignore momentarily the remaining data displayed in the graph. Students returned to their intuitive and informal sense of whether the coin was fair, rather than formally quantifying the difference between the observed and expected. At this point students struggled to construct and recognise mathematical meaning.

4.4.17 Detailed study For and Against – contextual sampling task

Part H of the detailed study explored sample size and measurement error in the context of a public opinion survey. Methodology for the task is presented in Section 3.7.9, and the worksheet is attached as Appendix D1, Part H.

In Part H students were presented with a fully functioning Fathom simulation workspace. The survey simulation was presented as iconic representations of human faces for the choice of either For or Against; this was designed to encourage students to imagine the survey as involving people. The simulation was used to explore the relationship between sample size and survey accuracy, but informally and without reference to the large population sample size model used in the classroom. Students were asked to determine the outcome of the simulated survey using the minimum sample size possible. The simulation was set within the range of 53–55% "For", but the students were unaware of this value.

Two students, E2611G and R2408I, examined the contextual task of a sample size for a large population survey from the perspective of the accuracy of the survey. The simulation was set to a default sample size of 10. Students E2611G and R2408I initially considered the relative frequency in a manner similar to comparing the relative heights of the bars in a column chart – that is, additive thinking. The students increased the sample size in small cautious increments of ten or twenty, perhaps taking a cue from physical simulations.

> E2611G – I had it on thirty [sample size] and it just wasn't big enough, so I put the sample up to 50 and it is still swapping between the two ….swapping between For and Against, sometimes it's 27 For and sometimes 28 Against.
>
> Researcher – …so if it is inconclusive, what would you do?
>
> R2408I – …put it up to 100.

> E2611G – now I think there is more people For I think, not dominating, just ahead.
>
> R2408I – yeh, every time I sample For is ahead. Sometimes it is ahead by a lot, other times ahead by two or four.

None of the students built-with their knowledge from the previous activity – the link between the two tasks was clearly too tenuous.

> Researcher – How accurate do you think your result is?
>
> E2611G – It depends on the population.
>
> R2408I – Couldn't just sample 100 people, it isn't enough.

Student R2408I's comment led the discussion to the sample size needed for Hobart's population of 200,000. The students were asked to discuss between themselves a suitable sample size. Student E2611G may have recalled the sample sizes used in the classroom activities and proposed a sample size of several thousand, and student R2408I proposed a very large sample size that many students had suggested much earlier in the pre-test (Section 4.2.4).

> E2611G – maybe 2000?
>
> R2408I – wouldn't you like it to be a third or something? Because a half would be too much.
>
> E2611G – so it's about 70,000, a bit less, over 60,000.
>
> Researcher – In the computer world if you use 5,000 or 50,000 it costs the same amount, but in an actual survey it will be very expensive to ask 50,000 people.
>
> E2611G – do you [speaking to the other student] want to go 5000 and I'll use a different number?
>
> Researcher – How about one of you do 2000, and you do 5000?

The students continued to use the frequency rather than a proportion, and it took a deliberate step by the researcher to shift students' focus from a frequency to the data expressed as percentage. Both students used the expression "dominant For," which suggests that they too were reporting the outcome of the survey and the difference in frequency, rather than the specific numerical value. In a sense the students "skipped-over" analysis of the data and went directly to the consequences of the information.

> E2611G – I have a dominant For.
>
> R2408I – I have a dominant For, but the difference was only 300.
>
> Researcher – What percentage did you get For?
>
> E2611G – 56% For.
>
> R2408I – I got 53%.

The researcher encouraged students to consider the accuracy of the survey result and the cost of conducting a survey. Student E2611G expressed accuracy as relative to fifty percent, rather than the scatter around the centre of the distribution of measures. Student R2408I immediately expressed the accuracy correctly using the same terminology used in the classroom activities.

> Researcher – … so you both tended to get a majority in Favour?
>
> E2611G – Yeh.
>
> Researcher – but you got the same answer [majority in favour] with half the cost.
>
> E2611G – Hmm.
>
> Researcher – [pause] do you have any feel for how accurate your result is? [pause] You have quoted the centre of your dots and you have a scatter around that result.
>
> E2611G – well, [pause] it seems accurate.
>
> Researcher – yeh, but how accurate? Try and put a number on it … perhaps think back to some earlier activities.
>
> E2611G – [pause] does 56 mean it is plus five [student is comparing with the 0.5 expected?]
>
> R2408I – mine would be plus or minus five.
>
> E2611G – so fifty percent is half, like half and half? Hers is Against is 44% and [For] is 56%, so is it 6%?

Students S1001J and T0612M took a different learning trajectory. The two students kept an informal tally of the proportion of simulations, in this instance 70%, in which the simulator gave a majority For, rather than the numeric proportion of the population that was in favour (the simulation was 53% For). The objective of the survey was to determine ultimately whether a majority was For or Against, so in a sense the students addressed the objective, but this approach did not allow the accuracy of the survey to be considered. Both of the students expressed the result in the negative, i.e., Against.

> S1001J – I have sampled it [re-sampled] maybe ten to twenty times, and it came up even maybe once and Against maybe twice.
>
> T0612M – I pressed it [ran the simulation] about thirty times…and it came up [Against] about six times.

Student S1001J, using a sample size of 80, continued to note the frequency rather than the proportion For. Despite the prompting of the researcher the student did not use the result expressed in percentage, apparently because the student was not confident in using percentage.

> Researcher – you had one result came down to 50%, how did it go the other way?
>
> S1001J – well, it was one below….it went to 39 and 41 [student is using a frequency count]
>
> Researcher – ohh ... okay … you are looking at the actual count [frequency]…can you think in terms of percentage?
>
> S1001J – I'm terrible at percentage!

The two students had observed that the simulation gave a result of For approximately in 70 or 80% of all simulation runs. The accuracy students considered referred to the proportion of simulations For, and both students preferred the use of the +/- notation.

> Researcher – How accurate do you think your results are? Well, you are saying the majority are For…
>
> S1001J – [very confidently and correctly] it is about plus or minus 20, 30….25% maybe. It doesn't go to the extremes a lot, I reckon +/- 25% is the max it can go up, but it doesn't seem to go down at all…I'd say +/- 25% , but plus 20%, -5%.
>
> Researcher – are you happier talking about a range here, or plus or minus?
>
> S1001J – I'd say plus or minus!
>
> Researcher – [speaking to other student] you thought it was 80%?
>
> T0612M – 80%, plus or minus 10%.

Students' thoughts on quantifying accuracy led to a discussion of meaningful or acceptable accuracy. To support students the researcher took a step back and discussed physical measurement of the purchase of a length of timber. This seemed to lead effectively to discussion of the accuracy of surveys and the interpretation of survey results.

> Researcher – [speaking to both students] so the first level of accuracy is to say the majority are For, and then the next step on is to say how accurate the result. Are you happy with an accuracy of 10%? If you wanted to buy a piece of timber it is pretty hard to ask for [precisely one metre]. You would be happy with how much either way?
>
> S1001J – I think I would be happy with two or three percent, I wouldn't want much less than a metre.
>
> Researcher – okay so you made a measurement here [referring back to sample] and you told me you were happy with a result that is +/- 20%?
>
> S1001J – yeh [tone suggests realises his error]
>
> Researcher – [speaks to other student] ….you wanted +/- 10%, which is 900mm rather than 1000mm. So you are happy with this situation but not that?
>
> S1001J – …so I'd be happy if it went above that is fine [when purchasing the timber].
>
> Researcher – …so you can cut a bit off at home [if the timber is too long]?

> S1001J – but if it goes minus it doesn't really matter [?]….minus 5% is probably the least I'd go.
>
> T0612M – [student applies the theme to the interpretation of the survey result to a majority "For"] with mine [re-sampling] I'd say comparing with 51% [a majority] because you only need 51% for a Yes.
>
> Researcher – Because it will flip the survey the other way?
>
> S1001J – Yeh!

Students had no difficulty with the concept that increasing the sample size increased the accuracy of the survey, but student S1001J demonstrated some sense, possibly as a consequence of the classroom study, that to achieve an accuracy of a 2% would require a sample size of several thousand.

> Researcher – If we wanted to be plus or minus two percent what sample size would we need to have?
>
> S1001J – …you would have to increase the sample size…you would have to put it up to [pause] ...
>
> Researcher – [pause] okay, I want you to increase the sample size until you think the answer will be within +/- 2%.
>
> S1001J – ….so changeable…like…I have gone right up to 3000 to see what happens.

At the larger sample size student S1001J's focus had evidently shifted from frequency to percentage.

> S1001J – …before when I was on [a sample size] of 80 …. I was still looking at the simple number, but now start looking at the percentage summary thing…..

Student S1001J's focus had shifted to the proportion of the population in favour, away from the earlier focus on the proportions of all the simulations that gave a majority For. The simulation result had tended to stabilise at approximately 55.3%, which was similar to the underlying value of 55%. The student used the convenient and sensible short-hand of truncating the data.

> Researcher – …are you still staying with a majority of 70%?
>
> S1001J – yeh, yeh, 70…. I don't know all of them seem to be 55% to 45 more….this one is 56 to 43 [truncating rather than rounding off]… with the bigger numbers you have to look at the percentages a bit more because the number [frequency] don't mean as much.
>
> T0612M – yeh, it's much easier with the percentages.
>
> S1001J – I have got it to plus or minus 4%.
>
> Researcher – …and your sample size?
>
> S1001J – 3000.

> Researcher – …and the number?
>
> S1001J – for the For? 55.3%. It's gotten more accurate than the 70% from the sample size of 80, it's still not right on.

Student T0612M continued to confound the proportion of times the survey simulation gave a result Against with the accuracy of the result. The researcher described the scatter, and helped students to identify that the centre of the distribution may have been significant.

> Researcher – what about you M?
>
> T0612M – uhhm. When you say are saying the percentage plus or minus 2%...are you saying like 2% of the time you sample it will be Against?
>
> Researcher – I picked a sample size and let's say 55, the next time I sampled at that same sample size it was 56, next time it was 55, then 52…it is mainly around 54.
>
> T0612M – so it was plus or minus?
>
> Researcher – plus or minus 3%, something like that…have you settled on a sample size?
>
> T0612M – …. at 1000 so far pretty good…it has gone to 4%.
>
> Researcher – so you'll go for 1000?
>
> T0612M – no, I'll go for 1500. [similar to the value in class]

Situated abstraction and Hershkowitz et al. (2001)'s three element model provide a framework for analysis of the students' responses. Students' development may have been highly individual, but some aspects were common. Students began by constructing knowledge by understanding the attributes, interpreting the graph, and observing frequencies or relative frequencies. It was either a large sample size or prompting from the researcher that shifted students' attention to recognise the data represented as proportions expressed as a percentage. Once students recognised the outcome as percentage, it was expressed initially as a pair, for example, "56 For and 44 Against," when in a dichotomous survey one value is redundant; i.e., "56% For" is sufficient. The process of construction and recognition occurred in tandem. The shift to percentage measurement also shifted the focus to the centre of the distribution of measures – an element of distributional thinking – rather than comparing single runs of the simulation representing additive or multiplicative thinking.

Four students used re-sampling initially to keep a tally of survey outcome that were either For (greater than half) or Against (a survey result less than half), rather than considering the underlying population proportion or the associated survey accuracy.

Students were building-with their knowledge and applying to the new context. Students had sensed that the outcome of the survey was the key aspect – a legitimate interpretation because it focused on the outcome of the survey as a majority being in favour. Student S1001J's sense of the accuracy of the survey was supported by first considering the practical significance of more familiar physical measurement. Practical significance differs from consideration of accuracy, because it considers meaningful measurement in a particular context. Students' understanding was, at that stage, too limited to allow the students to build-with effectively to the new context of the accuracy of surveys.

4.4.18 Colleague teacher interviews

The two colleague teachers were asked whether the topic of sample size and the sample size function were suitable topics for the two classes, and both agreed that the topic was indeed suitable. The colleague teacher of the male students thought that the more able students grasped the concepts and that the less capable students at least gained at the level of general conceptual understanding. The colleague teacher of the female students thought that the students needed more opportunities to practise skills. The question posed to the colleague teachers was whether the topic too ambitious.

> Not at all. This was a highly spread class. More able students would have grasped concept, but in this class only about half got a handle on it. Other students perhaps developed a feel for the concept [developing intuitions] and took away a general picture of how many to sample, even if they could not give a specific answer [of how large a sample size to take]. [Colleague teacher of male class]
>
> The content was very suitable. It simply needed more practice examples to reinforce the rule [large population sample model].There was a wide-range of abilities: some students got it all, many got the big picture, perhaps about 25% of students just did the mechanics and didn't get the big-picture. [Colleague teacher of female class]

The colleague teacher of the female students was asked whether the scope of the study should have been reduced by not including activities examining the fairness of the die or the contextual tasks, which would have allowed greater attention to the formal mathematical concepts.

> The work had to go all the way through to the Mt. Wellington cable-car. Students needed to be able to apply simulation to a real-life task, how true that result was. It had to be more than just a [computer] program. [Colleague teacher of female class]

The sample size model related sample size and accuracy. The colleague teacher of the male students was asked the value of consideration of accuracy.

> Whether you have the accuracy you want [alluding to the topic of the study] I hadn't spent as much time talking about the tolerance about the data as you did, but I now do that now in other areas such as measurement in grade 8.What you did reminded me of the importance because measurement is a great concrete way to start it and that gives them a good idea when they work with stats of the idea of tolerance and how everything is not exact and there is a spread that is acceptable in certain situations [Colleague teacher of male class]

In the study the sample size model was introduced relatively late in the study, and well after the context and purpose was established. Several students had commented to the researcher that they would rather see the function first, become comfortable with it and then use it later on. What approach would the colleague teachers use, or would different strategies be used for different students?

> I preferred the way you did it…because you are more likely to have a set of data and try and match a rule to it to help you in the future because the rules aren't just handed to you, they need to be tried to see whether they fit. I think they way you tackled it was a more realistic way to do it. The bottom line is I suspect that many students just want it handed to them [....] "give me a rule and I'll give you an answer," which is not really what we are after. [Colleague teacher of male class]
>
> I would introduce the concepts first, but perhaps it depends upon the circumstances [Colleague teacher of the female class]

4.4.19 Summary of findings for Research question 2

This sub-section examined the second of the three research questions: whether the large population sample model was accessible to Year 9 high school students. This was considered through students' expression of beliefs of sample size, use of a model of a single statistic of the fairness measure, use of the large sample size model, interpretation of survey accuracy, and application of the model to contextual tasks. These five aspects are addressed in turn.

Students' naïve beliefs of sample size in contextual tasks were established through the National and state election item given in the pre-test (Section 4.2.4). In the pre-test students favoured a far larger sample size than is used conventionally; a preference for using a smaller sample size for a smaller population; and the most favoured strategy of the strategies offered was a sample size of "10% of the population."

In the classroom and the detailed study the two broad types of tasks presented to students were either to justify a sample size given to them or provide a sample size independently. Students' responses were a subjective "large enough amount," a specific numeric sample size, or a sample size as a proportion of the population. The sample size used initially in the simulations appeared contextual with sample size for die and coin being initially similar to that used for physical simulation and the opinion surveys related to the population size.

Students' long-term development of sample size in contextual tasks was assessed using the same National and state election item re-presented as a follow-up test (Section 4.4.14) two months after the conclusion of the study. Assessment on this time-frame was designed to determine what sample sizes students would use outside of the classroom. Students' sustained long-term development of sample size was modest. Students' beliefs of sample size had changed little and they persisted with use of a larger sample size than is used conventionally, but students did, however, tend to adopt a more consistent sample size strategy for the two different populations of the national and state opinion polls. The most remarkable change was students' shift from a strategy of 10% of the population to an inappropriately large specific numerical sample size.

The study examined two types of mathematical models. The first type of model were the two single statistics of the fairness measures, The first single statistic model was the fairness measure which formally quantified the fairness of three dice and the second was an extension to the %fairness measure, Students examined the formal mathematics supporting the fairness measure through a homework item where students worked independently. The students who submitted a response (84% of girls, 62% of the boys) had little difficulty completing the tasks and this indicated that, for these students, the fairness measure statistic was within their conceptual and mathematical grasp. In a subsequent extension to a %fairness measure, the measure was not conceptually understood by student generally, and students misapplied the principles in their consideration the measure of a proportion of heads at the two sample sizes of 50 and 500 tosses of a coin.

The second, and principal model of the study, was the large population sample size model = $\pm 1/\sqrt{n}$, which related the margin of error, e, to the sample size, n. In the pre-test 76% of the boys and 64% of the girls substituted a simple integer into the model successfully. Students first examined the accuracy of the model using a process of statistical enquiry, a frequentist approach, and a Fathom simulation of a multiple coin toss. In their formal consideration of the model 67% (23/34) of the female and 38% (8/21) of the male students provided multistructural responses or above in the examination of the data, and students generally concluded that the sample model was accurate. Students' mathematical understanding of the sample size model was explored through two items on the post-study assessment. The boys' performance was modestly superior to the girls' on both items, but only 30% of the students responded correctly.

The study applied the sample size model to the contextual task of a public opinion survey for the Mt. Wellington cable car. Students' naïve beliefs of sample size on this item were consistent with the students' responses to the earlier National and state election survey and were predominantly related to a sample size as a proportion of the population (Section 4.4.8). In the post-study assessment, which is the time-frame used in conventional school assessment and the one most familiar to practising teachers, both female and male students demonstrated development of understanding of sample size. In the post-study assessment 48% (17/36) of the females provided a relational response to the item. The level of responses provided by the males was not as sophisticated and one only male (6%) provided a relational response, but 56% (10/18) provided a multistructural response or higher. Students appeared to be able to hold simultaneously the contradiction of a formal mathematical and a subjective sample size strategy.

Students' use of the sample size model did not show students' conceptual beliefs and conceptual understanding of accuracy and error associated with sampling, so these beliefs were examined somewhat separately. Students brought to the study their own intuitive sense. Students compared their intuitions of the proportions of heads of a class set of 50 tosses of a coin established in the pre-test with the data of a Fathom simulation. In the pre-test male students tended to believe generally that the proportion of heads occurred in a narrower range than can be justified by chance. The intuitions of the female students were both more cautious and accurate than the male students, but the female responses differed little from responses chosen randomly (Section 4.4.7). When students compared their intuitions with the Fathom simulation students tended to be satisfied with their predictions, but this confidence could not be justified mathematically. Only when students' predictions were grossly inaccurate did they recognise that their intuition was incorrect.

In the detailed study examining the simulation of the public opinion survey students' strategies were most commonly to determine the relative outcome of a series of simulations, or, when the sample size was increased, to seek a stable numerical result. Students needed to be prompted by the researcher to quantify the error by calculating the difference between the observed and expected, and in this calculation students tended initially to misplace the decimal point. For one pair of students relating the acceptable accuracy of physical measurement, in that instance of a length of timber, provided an effective means to consider the accuracy of surveys. Students' notions of accuracy examined in the post-study questionnaire suggested only that the principle of increased sample size increased accuracy was understood.

In the detailed study's large population contextual task of the For or Against surveys many students proposed initially the same large sample sizes used incorrectly in the study pre-test. No student in the detailed study transferred the learning of the coin simulation to contextual tasks readily, and students required the support of the researcher to shift their focus to consider the accuracy of the survey. Students used the survey simulation initially to determine whether the survey predicted a majority for or against, rather than seeking to determine the underlying numerical survey result or the associated accuracy, which suggested students' abilities to extend and build-with and apply the model to new tasks were limited.

The detailed study included an examination of students' learning pathway of conceptual understanding of sample size concepts. Hershkowitz et al.'s (2001) three element model of constructing, recognising, and building-with was used to trace students' development of thinking as they examined data generated by a simulation of a coin-toss and a simulation of a contextual survey task. In the simulation of a coin toss, students demonstrated the first element of constructing where they had no difficulty in noticing informally that the cumulative proportion of heads tended to approach the expected value as the sample size increased. The second element, recognising, was supported by repeated simulation runs appearing to encourage students to look for patterns and develop proto-theories. To demonstrate the third element, building-with, students required the direction of the researcher to shift to formal measurement of the difference between observed and expected.

The second research question considered whether the sample size model provided an accessible method for high school students to determine sample size. The model was used in a study that introduced Fathom and re-sampling methods, and which sought to support students' development of intuitive understanding of sample size and survey accuracy. Students' development of understanding of sample size was modest, but some evidence exists to show students sense of sample in contextual surveys had changed to, away from a sample size of 10% of the population to smaller, but inappropriately large sample sizes in contextual surveys, and to larger sample sizes when using virtual simulation. Students' development of an intuitive sense of accuracy of surveys was limited, and the study provided an introduction to the concept only. Students found the calculation involving accuracy of decimals less than one difficult initially, but opportunities for practice reduced this difficulty. Students' mathematical understanding of the fairness measure suggested the fairness measure was within their grasp, but their mathematical understanding and application of the more complex sample size model can be considered introductory only.

4.5 Research question 3: Fathom re-sampling as a tool for high school

4.5.1 Introduction

This sub-section provides the data for the third of the three research questions and evaluates Fathom virtual simulation and re-sampling as an effective learning opportunity for high school. The data from the classroom work samples and the detailed study were re-examined and three aspects, peculiar to Fathom and re-sampling, were chosen for analysis: (a) re-sampling terminology, (b) measures dot plots, and (c) students' relationship with Fathom. The three aspects of the evaluation tasks first presented in Table 3.9 as the methodologies are re-presented in Table 4.34 as a summary of the data collected for the tasks.

Table 4.34.
Evaluation Tasks for Research Question 3: Fathom Re-sampling as a Tool for High School

Aspect of research question	Classroom study			Detailed study
	Initial task	Developmental task	Final task or assessment	
Re-sampling terminology (Section 4.5.2)	Develop the Fairness Measure (Sections 4.3.3 & 4.5.2.1)	Coin measures homework – Part 1 (Section 4.5.2.2)	50 students in a Year 9 maths class (Section 4.5.2.3) Post-study questionnaire (Section 4.5.2.4 Item 6)	Part F: Cumulative proportion of heads and sample size graph (Section 4.5.2.5)
Measures dot plot (Section 4.5.3)	Female race-times, Pre-test Q.5 (Section 4.5.3.1)	Compare three dice using GICS (Sections 4.3.6 & 4.5.3.2) The effect of sample size on the fairness measure (Sections 4.4.3 & 4.5.3.3). Coin measures homework – Part 2 (Section 4.5.3.3)	Badly biased coin (Section 4.5.3.4) Mixed up measures (Section 4.5.3.5) Post-study questionnaire (Section 4.5.3.6 Item 1 & 8)	Part F: Cumulative proportion of heads and sample size graph (Section 4.4.14) Part G: Measures dot plots at sample sizes of 50 & 500 (Section 4.5.3.7)
Students' relationship with Fathom (Section 4.5.4)	New York marathon – introduction to Fathom (Sections 4.2.6 & 4.5.4.1)	Fathom virtual die (Sections 4.3.5 & 4.5.4.1) The effect of sample size on the fairness measure (Sections 4.4.3 & 4.5.4.1)	Fathom basic skills test (Section 4.5.4.1) Students' perception of Fathom (Section 4.5.4.3) Post-study questionnaire (Section 4.5.4.4 Item 3, 4, 18 & 29)	Part A: Die one face only, or six faces presented simultaneously (Section 4.5.4.6) Part B: Coin side only or both sides presented simultaneously (Section 4.5.4.6) Part C: The effect of sample size on preference for data representation (Sections 4.5.4.2 & 4.5.4.6) Part D: Three potentially biased virtual coins (Section 4.5.4.2) Part E: Graphs and the Law of Large Numbers (Section 4.5.4.2 & 4.5.4.6)

4.5.2 Re-sampling terminology

This sub-section examines students' use of the two key expressions: "sample size used to calculate the measure" and the "number of measures collected." The two expressions provided indicators of students' understanding of the principles used in re-sampling.

In conventional sampling a sample of size *n* is used and one only statistic or measure, e.g., mean, is calculated. Re-sampling repeats sampling from the population with the sample size *n*, and more than one measure is collected. The hypothesis was that students would confound a sample and a set of measures, and to make the distinction clear the expressions "sample size used to calculate the measure" and "number of measures collected" were used. The expressions, although cumbersome, were designed to minimise students' confounding the two different ideas.

4.5.2.1 Develop the fairness measure

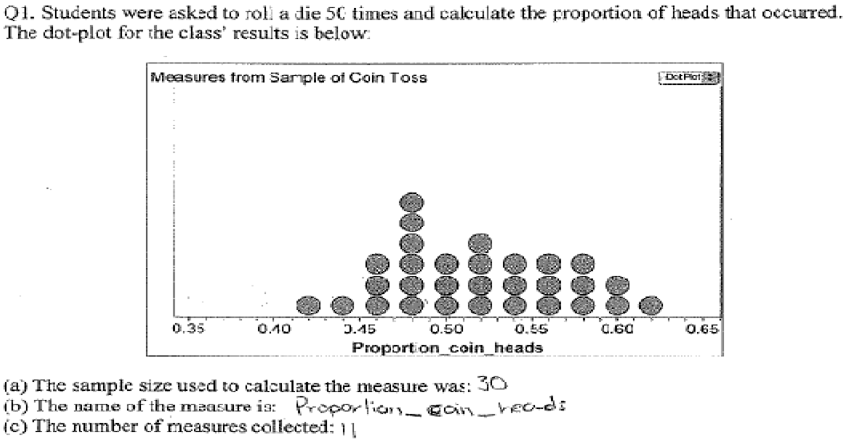
Students' first exposure to the expressions "sample size used to calculate the fairness measure" and "the number of measures collected" was the fairness measures activity (Section 4.3). Students had no apparent difficulty in calculating the fairness measure for each of the three die and contributing their results to the fairness measure dot plots (Figure 4.5). The number of measures collected roughly equated to the number of students present in the class, but there was a sense in the class discussion that the number of measures was of less importance than the sample size used – a different class would have collected a different number of measures. The distinction between sample size used to calculate the measure and the number of measures collected, at the conclusion of the examination of the fairness of the Fathom die, seemed clear.

4.5.2.2 Coin measures 50 & 500 tosses of a coin homework – Part 1

The item assessed students' ability to use "sample size used to calculate a measure" and "the number of measure collected" in, what was at that stage, the unfamiliar context of a coin toss. The item asked students a series of three questions: (a) identify the sample size, (b) provide a suitable name for the measure, and (c) state the number of measures collected. Of the two classes 48% of boys and 64% of the girls submitted a response to the homework item. Methodology for the item is presented in Section 3.3.4.2.

Students had little difficulty (90% of male and 90.5% of female students correct) in identifying correctly the sample size used. Identifying the "number of measures collected" yielded a mixed response with all male students correct and 76.2% of female students correct. Of the five female students who were incorrect one student confused the sample size used to calculate the measure with number of measures collected; one student multiplied "sample size" and "number of measures" and calculated the total number times the virtual coin was tossed (e.g., 50 * 30 = 1500); and three students provided a response of eleven, which was apparently the number of columns in the measures dot-plot – the students appeared to have counted the number of the different proportions of heads that occurred, i.e., 0.42, 0.44, 0.46, 0.48 ... 0.62, rather than the actual number of measures collected (Figure 4.16). The most challenging item for all students, with 50% of male and 66.7% of female students correct, was providing a meaningful name for the measure, such as "proportion of heads," even though an attribute title "Proportion_coin_heads" was displayed on the graph axis (Figure 4.18). Of the five boys who gave an incorrect response four chose the graph title of "Measures from sample of a coin toss" and one boy gave the response "expected proportion of heads." This suggested that a significant proportion of students did not have a global understanding of the representation of the data set.

Student N3110T provided an illustrative example of a student who did not have a strong understanding of measures. The student identified correctly the sample size used to calculate the measure, provided a name for the measure, but incorrectly gave the number of measures collected as eleven rather than the correct 30 (Figure 4.18). Several students, all females, gave this same incorrect response.



Q1. Students were asked to roll a die 50 times and calculate the proportion of heads that occurred.
The dot-plot for the class' results is below:

Measures from Sample of Coin Toss
DotPlot
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Proportion_coin_heads

(a) The sample size used to calculate the measure was: 30
(b) The name of the measure is: Proportion_ coin_ heads
(c) The number of measures collected: 11

Figure 4.18. Student N3110T response to 50 & 500 tosses of a coin homework.

4.5.2.3 50 students in a Year 9 maths class

In the post-study assessment Q.1 students were set a similar task to the previous item and asked the same three questions given in the Coin measures 50 & 500 tosses of a coin homework (Section 4.5.2.2): (a) identify the sample size used to calculate a measure, (b) provide a meaningful name for the measure, and (c) state the number of measures collected. In contrast with the first homework item, the post-study item included responses from essentially all students participating in the study. Methodology for this item is presented in Section 3.3.5.1, and the results for the three parts of the item are discussed in sequence.

Table 4.35 Q. 1 (a) shows that 72% of boys and 65% of girls identified the sample size used to calculate the measure correctly. The most common error was students confounding the sample size with the number of students who collected data. In conventional sampling familiar to the students this would indeed be the sample size, but in re-sampling this was the number of measures.

In response to Q. 1 (b) students, particularly boys, continued to find providing a meaningful title to the measure difficult with only 39% of boys and 60% of girls able to provide an appropriate title; this was a level of response similar to the "50 & 500 tosses of a coin" item above. An example of an appropriate title was "proportion of heads" (Student D1312Z) and an example of an inappropriate title was "the fairness measure" (Student G0709A).

In response to Q.1 (c) the boys again had little difficulty identifying the number of measures collected, but only 40% of girls gave the correct response. The most common incorrect response (45.7% of the girls who responded) was to multiply the sample size by the number of measures used. The explanation for this response lies with the whole-class discussion in the girls' class. The researcher attempted to promote the speed and efficiency of simulation and calculated the (very large) number of times the virtual coin was tossed by multiplying the sample size with the number of measures, but this appeared to confuse the girls, so the topic was not discussed subsequently in the boys' class. The girls who responded incorrectly may have felt obliged, when given two numbers, to perform a mathematical operation.

Table 4.35.

Post-study Assessment Q. 1 Students' Understanding of "Sample Size Used to Calculate Measure" and "Number of Measures Collected."

Post-study item	Boys		Girls	
	N=18	%	N=35	%
Q. 1(a) Identifies sample size				
Correct	13	72 %	23	65 %
Confounds with measures	2	11 %	8	23 %
Confounds with total data collected	3	17 %	1	3 %
Incorrect, unclassified	0	0 %	2	6 %
No response	0	0 %	1	3 %
Total	18	100 %	35	100 %
Q. 1(b) Name for Measure				
Appropriate title	7	39 %	21	60 %
Ambiguous title	8	44 %	9	26 %
No response	3	17 %	5	14 %
Total	18	100 %	35	100 %
Q. 1(c) Number of Measures collected				
Correct	15	83 %	14	40 %
Incorrect, 1000, multiplies sample size and number of measures	0	0 %	16	45 %
Incorrect, gives sample size	0	0 %	2	6 %
Incorrect, unclassified	3	17 %	0	0 %
No response	0	0 %	3	9 %
Total	18	100 %	35	100 %

In the classroom study students were provided with a developmental pathway. This pathway included re-defining the terms sample and measure, the first use of the two expressions "sample size used to calculate the measure" and "the number of measures collected" in a highly supported investigation of the fairness of a die. Subsequently, students worked progressively more independently in a series of tasks of the 50 & 500 tosses of a coin homework item, the Fathom coin activities examining sample size, and concluded with the post-study assessment. At the conclusion of the study many students were unable to provide a meaningful title for the measure, which suggested these students did not understand the underlying data set. Approximately 2/3rds of both male and female students identified the sample size correctly. More than twice the proportion of male students (83%) than female students (43%) identified the number of measures collected correctly. The substantial difference in the male and female students' abilities to identify the measures correctly could be explained by the researcher's calculation of the total number of times the virtual coin was tossed in the girls' class.

4.5.2.4 Students' post-study questionnaire items

The post-study questionnaire Item 6 examined students' self-assessed confidence in the distinction between sample size and the number of measures collected. Table 4.36 shows that female students were not confident of the distinction between sample size and the numbers of measures calculated with only slightly more than half (54%) saying they did not feel confused. Students' self-assessed confidence may be at odds with their ability to actually make the distinction between sample and measure.

Table 4.36.
Post-study Students' Questionnaire Sample Size and Measures Collected Items

Post-study questionnaire item		Disagree or Strongly disagree %	Maybe or neutral %	Agree or Strongly agree
6. I was confused by "sample size" and "number of measures calculated"	M (n = 18)	72 %	22 %	6 %
	F (n = 35)	54 %	23 %	23 %

4.5.2.5 Detailed study

The detailed study Part F provided an opportunity to examine the two expressions of "sample size used to calculate a measure" and "the number of measures collected" with students in more depth. Students were asked to reflect back to a time before the study to describe their natural language sense of the word measure; students variously described measure as the act of measurement, or measurement of a familiar physical parameter.

> To measure something physical…a verb. [Student N2610H]
>
> Before Fathom it was volume or something. [Student R2408I]
>
> Measurement in centimetres. [Student T0608I]

Student pairs were then given tasks designed to examine the distinction between sample size and the number of measures collected. Students CN2701B and Y1504L considered a scenario where 25 students each tossed a coin 50 times and calculated the proportion of heads, tasks students had experienced in the classroom study some six weeks earlier.

> Researcher – In your class I've asked each person to flip a coin 50 times and work out the proportion of heads. What is our sample size, and how many measures would we collect?
>
> N2701B – [long pause 4 secs] Oh I think I'd go 25, 25 as a class as a sample size…oh…[still considering]
>
> Y1504L – 25 measures…because …

> N2701B – yeh.
>
> Y1504L – …because there is 25 people doing it…
>
> N2701B – yeh…hang on…oh… the question?
>
> Y1504L – how many times were they rolling it?
>
> Researcher – 50.
>
> Y1504L – 50, then sample size is 50.
>
> Researcher – yeh, sample size is 50…ok, ok…

…and subsequently considering rolling a die:

> Researcher – what did each one of those dots [on the dot plots] represent? You made some good points….
>
> Y1504L – [very confidently] one measure [correct]
>
> Researcher: – one measure…was that one roll of the die?
>
> Y1504L – it was whatever the sample size was [correct]

A second student pair, N2610H and R1706D, considered re-sampling in the context of a public opinion survey. This part of the discussion was dominated by a generally capable and confident student N2610H, but the student was not, however, confident of the distinction between sample size and measures. The researcher prompted the students by asking their understanding of the meaning of the measure in the context of an opinion survey, which may have been a more challenging context than coin tossing, because students may have been accustomed to re-sampling activities where a large number of measures were collected. N2610H began by constructing mathematical knowledge.

> N2610H – Sample size is 50 because you have asked 50 people.
>
> Researcher – and what is the measure?
>
> N2610H – I'd probably go one per person.
>
> R1706D – Because they only have one opinion…
>
> N2610H – Well there is only one question in the …
>
> R1706D – …and there can only be one answer…
>
> N2610H – Not two or three questions otherwise there would be two or three measures…with the dice thing 20 times that is one measure…one sample with 20 measures [student has incorrectly reversed the concept]

Student N2610H, appearing to be uncertain, changed to describing measures in terms of a coin toss. The student's thinking appeared to be moving between constructing and recognising mathematical knowledge to assemble a coherent mathematical structure.

> N2610H – …exactly what I was saying because it didn't sound quite right…if you flip a coin 50 times…that's 50 measures…no, no, sample size is 50 for that

> Researcher – sample size is 50?
>
> N2610H – sample size is 50 for that...
>
> R1706D – you are losing yourself!
>
> N2610H – yes…I'm confusing myself …that bit on the sheet about confusing has done just that cause I'm confused but 50 is …
>
> Researcher – [pause] what are you trying to work out at the end of the 50?
>
> N2610H – Basically it's the sample size is 50…basically the sample size is 50 when you flip the coin 50 times.
>
> R1706D – Yes.
>
> N2610H – and the number of measures how many different times you did that, the 50 flips [correct].

The student then returned to the original context of an opinion survey. The students used the novel approach of a sample size of 200 split into four groups of 50 allowing four measures to be calculated. The student confidently interpreted the measure correctly.

> N2610H – if you are measuring 200 times and split it into four groups of 50 and you used each one [of the four] as a separate thing, then that would be four measures.
>
> Researcher – oh ok… what is the actual measure you are interested in? Is it proportion of heads…what are you trying to calculate at the end of your 50?
>
> R1706D – Against or For.
>
> Researcher – oh ok.
>
> N2610H – you'd use For probably…proportion of people For.

The students felt confident in their understanding of the two key terms, but this confidence may have been at variance with their actual understanding. The students thought that activities that displayed the dot plots as a class poster supported their understanding, a view also expressed by the colleague teacher of that class. The two students reflected on the classroom study:

> R2408I – I got a bit confused between measures and sample size, and which one they actually went with, especially in the homework bit [Section 4.4.4], but in the class I thought it was clear.
>
> E2611G –Yeah, yeah, the dot points on the wall made it clear.
>
> R2408I – That [dot plot displayed in the classroom] made it easy to understand.

In another exchange two male students also displayed similar confidence, at least by the end of the study, regarding their understanding of measures. One student used the word "easier" that conveyed a degree of hesitancy.

> S1001J – Sometimes you have to stop and think, but most of the time I was fine with it.
>
> T0612M – I got pretty confused with the difference between them …. in the beginning.
>
> S1001J – Yeah, in the beginning, but in the end it was easier.

Students experienced difficulty with verbal tasks that used the two expressions, but graphical representations of measures were less troublesome; students needed to count the number of measures only, but the sample size used was less obvious. In the following example a student independently calculated the measures collected for one sample size. The student's apparent ease of use here was similar to their apparent ease of use of the dot measures dot plot of the fairness measure.

> Researcher – how many [measures] did you collect?
>
> R2408I – 24 …32 less 56 [determines the number of measures by mentally calculating the difference between the total number of measures and the measures for the other sample size]
>
> E2611G – [laughs] like 24 or something…is that right?
>
> Researcher – […] it is just making sense of the information.
>
> E2611G – yeh
>
> R2408I – 24.

4.5.2.6 Summary of findings for re-sampling terminology

By the conclusion of the classroom study many students remained confused by the two terms "sample size used to calculate a measure" and "number of measures collected," and students expressed lack of confidence in making the distinction. Students apparently found the terms comprehensible initially in the examination of the fairness of the dice, but subsequent application to other contexts difficult. Some students' confusion may have been compounded by their apparent lack of understanding of the underlying set of re-sampled data.

Students' understanding of a measure within a particular context, shown by such as the examination of the fairness of the dice or by giving an example, suggests that students had an exemplaric and low level understanding where the word is described by example (Meyer, 2009, p. 910), rather than by definition. Their difficulties arose when the principle of measures was generalised to contextual tasks, such as the short-worded problems given in assessment tasks, which suggested that students did not have a definitional sense of the terms. Students found the distinction clearer when the data were presented as measures dot plots such as were used in the study of the fairness of the dice and subsequently in the detailed study.

This study was students' first formal experience of re-sampling and the two expressions. The students faced a number of challenges of developing a new or changed meaning and definition of sample and measure terms, making a distinction between two new concepts, and carrying the two terms simultaneously – a clear increase in abstraction. Students needed to adapt their natural language use of the terms sample and measure to the alternate and formal use for re-sampling. The word measure was a natural one to students, but as a verb, not as the noun used in the study. The changed definitions occurred within an unfamiliar context of re-sampling. Students' self-assessed confidence as expressed in the questionnaire and in the detailed study may be at odds with their ability to use the two expressions properly.

4.5.3 Measures dot plots

This sub-section examines students' development of their ability to interpret measures dot plots. The measures dot plot is a dot plot of a collection of measures generated during re-sampling, for example, a collection of proportion of heads. The measures dot plot may promote the two key principles of re-sampling that as the sample size increases the spread of the measures decreases and that the centre of the distribution of measures approaches the expected value.

Eight activities and assessment items provided evidence of students' development and use of measures dot plots. The dot plot format was first introduced (or re-introduced) to students in the pre-test Q. 5 as a data analysis task of Female race-times. Measures dot plots were first introduced to the students in the examination of the fairness of dice, which was formally assessed in the task that compared the fairness of dice using GICS (Section 4.3.6). The three tasks "The effect of sample size on the Fairness Measure", the Coin measures "50 & 500 tosses of a coin" given first as a homework item, and subsequently as a classroom simulation as "Fathom virtual 50 and 500 tosses of a coin" were classroom developmental tasks. Students' use of the measures dot plots was formally assessed on the post-study assessment with the two items "Badly biased coin" and the item "Mixed up measures." This evidence is supported by items from the post-study student questionnaire, extracts from the detailed study and the interview with the colleague teachers.

4.5.3.1 Female race-times (Pre-test Q. 5)

The dot plot was introduced to students as a pre-test item to establish students' ability to interpret the dot plot representation of a simple and experiential data set. Table 4.37 shows that 72% of male and 88% of female students provided a multistructural response or higher; students had little difficulty interpreting correctly a dot plot of a familiar data set. The most challenging aspect for students was identifying the centre of the data item formally as the mean, median or mode, but a formal definition was not essential to demonstrate students' intuitive sense of centre. Examples of unistructural, multistructural, and relational students' responses are attached in Appendix G.1.

Table 4.37.
SOLO Evaluation of Students' Responses Female Race-time Item

SOLO	Boys		Girls		
level	n=18	%	n=33	%	Criteria or exemplars
NR	0	0%	2	6%	No response
U	5	28%	2	6%	One or two criteria only (see below)
M	11	61%	26	79%	Partial understanding, at least three criteria met
R	2	11%	3	9%	Demonstrates sound and integrated understanding of five criteria: – Correctly identifies fastest time – Provides an appropriate range of "most" – Correctly orientates graph (fastest race-time being lowest value) – Locates centre of data – Identifies centre correctly as median, mean, or mode.
Total	18	100%	33	100%	

4.5.3.2 Compare three dice using GICS

The first formal assessment of students' ability to interpret a measures dot plot was the task "Compare three dice using GICS." (Section 4.3.6). This was a very complex task, but the levels of response, particularly from the female students, were high and Table 4.5 shows that 57% of male students and 97% of female students provided multistructural or higher responses. The task was complex because the data were of a derived statistic, three measures dot plots were presented and analysed simultaneously, and the students worked independently under traditional examination conditions. The task was supported by the GICS framework, which was designed to provide a checklist of all aspects of the data, notes taken during class discussion, and the measures dot plot displayed at the front of the class.

4.5.3.3 Sample size, the fairness measure, and 50 & 500 tosses of a coin

The two developmental tasks of "The effect of sample size on the Fairness measure" (Section 4.4.3 and Table 4.17) and the "Coin measures 50 & 500 toss of a coin – Part 2" (Section 4.4.5 and Table 4.18) are examined as two companion tasks because many students transferred their learning of the first task to the second task incorrectly.

The first task, "The effect of sample size on the fairness measure," was a highly supported classroom activity where students used a Fathom simulation to generate %fairness measures at sample sizes of 30, 300, and 3000 rolls of the Fathom die. Students contributed their own data to a class set, participated in whole-class discussion analysing the data, and reproduced the dot plots displayed in the classroom. For the second and companion task, "50 & 500 tosses of a coin" homework item, students worked independently applying their developing knowledge of measures dot plots to what was at that stage the unfamiliar data of a collection of proportion of heads. Methodologies for the tasks are presented in Sections 3.3.4.1 and 3.3.4.2. The two tasks are examined in sequence.

To complete the first task – the classroom activity examining the effect of sample size on the %fairness measure – successfully students replicated the series of three measures dot plots displayed in the classroom, included titles, axis labels and measures of centre, interpolated to two the sample sizes of 900 and 1600, and described verbally the effect of sample size on the measure. Three student exemplars are presented.

Student Y1504L provided a unistructural response (Figure 4.19) that reproduced the dot plots and included a measure of centre, but the dot plots did not include scales or axis titles, and the student did not interpolate to other samples sizes. The student described the data spread as "more grouped," but did not indicate whether this occurred when the sample size increased or decreased.

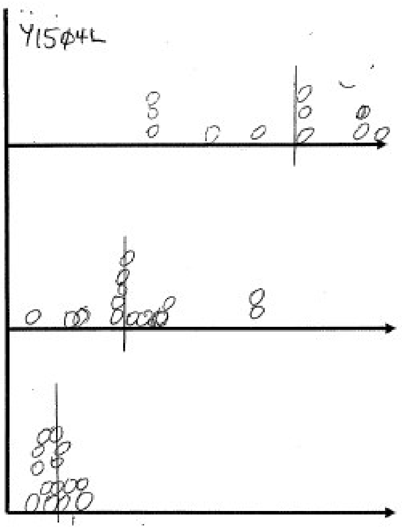


Figure 4.19. Student Y1504L's unistructural response to the effect of sample size on the %fairness measure.

Student N0909C provided a multistructural response (Figure 4.20). The student constructed a meaningful graph that included scales, labels, and measures of centre. The student interpolated to sample sizes of 900 and 1600, but provided inappropriately accurate values of the %fairness measure ("8.9 and "7.2") and did not include the values on the graph axes.

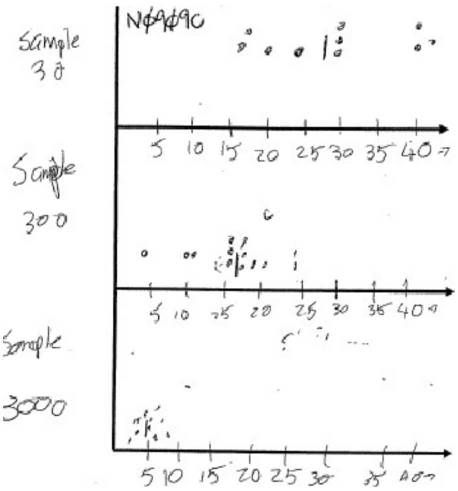


Figure 4.20. Student N0909C's multistructural response to the effect of sample size on the %fairness measure.

The student's verbal response suggested a confused understanding of fairness.

> As the sample size gets smaller the centre gets smaller. As the sample size gets bigger the spread is smaller. At a sample size of 900 the centre [would be at] 8.9 and at 1600 the centre [would be at] 7.2. The die is fairer as the sample size increases. [Student N0909C]

Student G0709A provided a relational response. The student's carefully crafted graph, presented earlier as Figure 4.6, included scales, labels, and measures of centre, and the student interpolated correctly to the two sample sizes of 900 and 1600. The student described the centre of the data as "closer to zero" as the sample size increased. The behaviour of the spread of the data was described as "the larger the sample size the more gathered the measures."

To complete the second and companion task of the 50 & 500 tosses of a coin homework – Part 2 item students sketched the proportion of heads of a series in 500 tosses using a 50 coin toss as a template. Student E3011L (Figure 4.21) gave a correct response and sketched a measures dot plot with the distribution centred at a proportion of heads close to 0.5 and a range of the distribution narrower than for the sample size of 50.

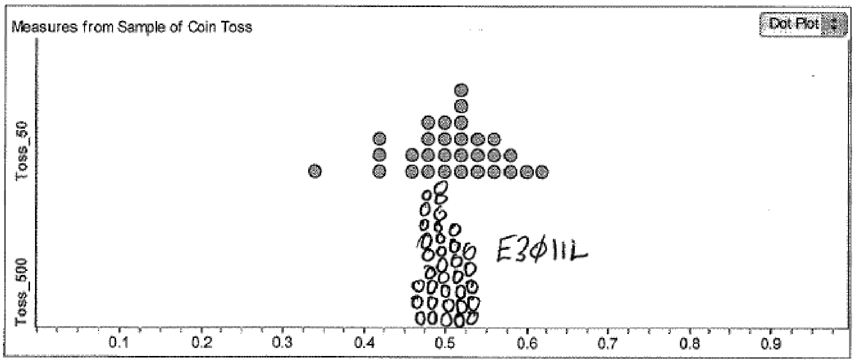


Figure 4.21. Student E3011L 50 & 500 tosses of a coin homework item showing distribution centred correctly at a proportion of heads of 0.5 and a distribution narrower at a larger sample size.

Student I0812A provided a response illustrative of that given by 25% of male and 57% of female students. The student recognised that the data spread decreased as the sample size increased, but appeared to have taken a cue from the effect of sample size on the fairness measure activity and placed the centre of the data substantially to the left. This response implies that the coin became biased as the sample size was increased, and suggested that students did not fully understand the meaning of the graph (Figure 4.22).

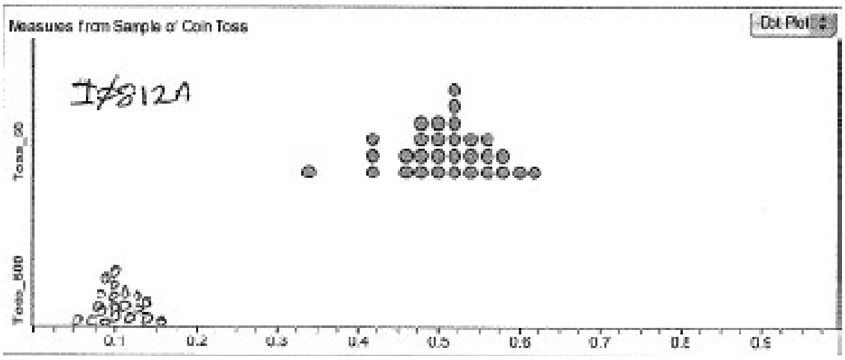


Figure 4.22. Student I0812A 50 & 500 tosses of a coin homework item showing distribution centred incorrectly at a proportion of heads not approximately equal to 0.5.

4.5.3.4 Badly biased coin

By the post-study assessment a majority of students had resolved the misconception demonstrated in the previous sub-section. The "Badly biased coin" task asked students to sketch a measures dot plot of a coin biased towards heads. To complete the task successfully students sketched a data distribution centred at a proportion of heads between 0.5 and 1.0. Methodology for the item is provided in Section 3.3.5.5, and the item is attached as Appendix A.15, Q. 4. The results, presented in Table 4.38, show that male (72.1%) and female (74.2%) students gave similar levels of correct multistructural or higher responses. Many students sketched a graph containing data of an unrealistically biased die.

Table 4.38.
Students' Responses to the Badly Biased Coin Post-study assessment Q. 4

Student's response	Male students		Female students	
	No.	%	No.	%
No response	2	11.1 %	5	14.3 %
Prestructural (incorrect)	1	5.6 %	2	5.7 %
Unistructural (partially correct)	2	11.1 %	2	5.7 %
Multistructural (correct)	12	66.7 %	24	68.6 %
Relational (correct)	1	5.6%	1	2.9 %
Total	18	100.0 %	35	100.0 %

Student R1207L provided an incorrect and prestructural response. The student sketched a measures dot plot, but the coin was so subtly biased as to be indistinguishable from a fair coin. The student may simply have reproduced a measures dot plot that was displayed in class. The scale was drawn with only sufficient information provided to identify the centre of the distribution.

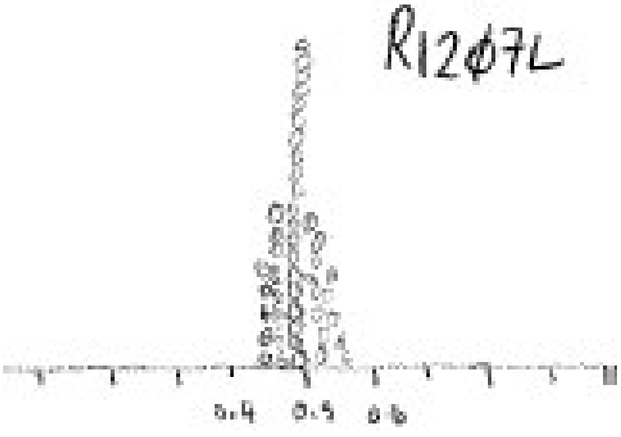


Figure 4.23. Student R1207L's prestructural response to the badly biased coin item showing negligible bias.

Student S1001J provided a multistructural response (Figure 4.24). The student had chosen a cumulative proportional of heads graph rather than the measures dot plot that was envisaged by the researcher. The response is not incorrect because the student had demonstrated clearly that the coin was biased towards heads. The plot appeared to equal a value of one, but this cannot occur if a tail had occurred at some point in the coin toss. One other student provided a response of this form.

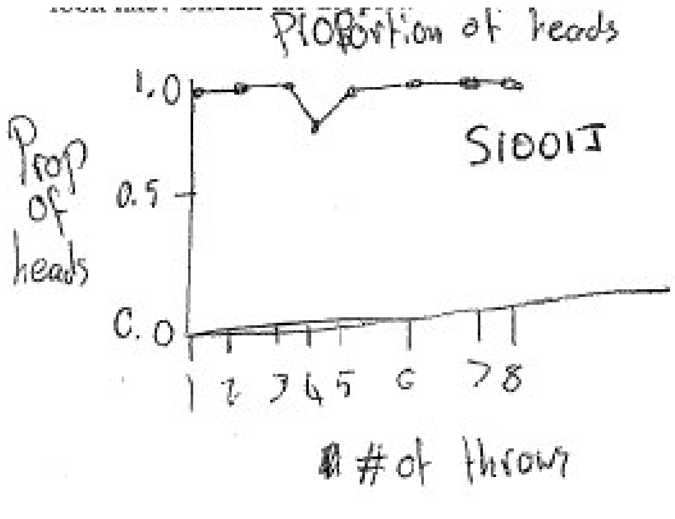


Figure 4.24. Student S1001J's multistructural response to the badly biased coin item.

Student N0106D provided a multistructural response (Figure 4.25) that correctly centred the distribution to a proportion of heads substantially larger than 0.5, but the distribution also included data greater than 1.0, which is impossible. The notation, which includes 010 to denote 1.0, suggests that the student was troubled by decimal place.

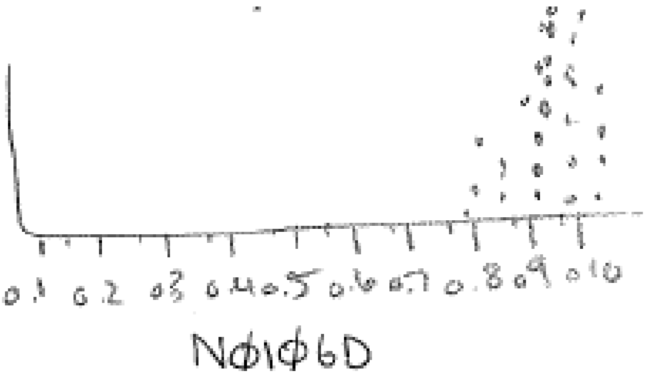
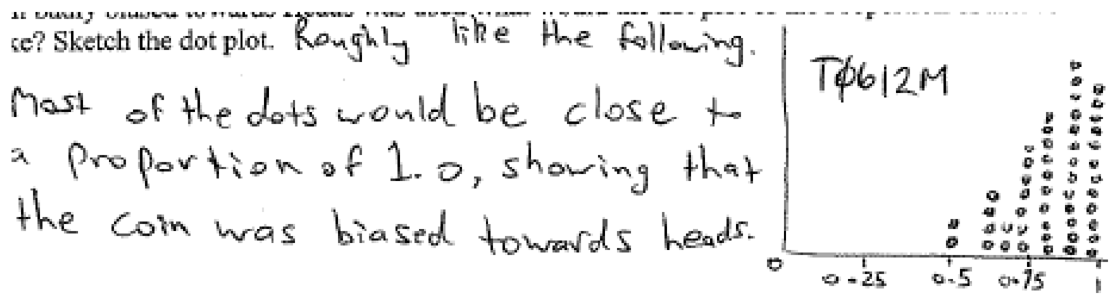


Figure 4.25. Student N0106D's multistructural response to the badly biased coin item.

Student T0612M provided a relational response. The student centred the distribution unambiguously above 0.5 and provided a correct brief verbal explanation of the dot plot.


ce? Sketch the dot plot. Roughly like the following. Most of the dots would be close to a proportion of 1.0, showing that the coin was biased towards heads.

T0612M
0
0.25
0.5
0.75
1

Figure 4.26. Student T0612M's relational response to the badly biased coin item.

4.5.3.5 Mixed-up measures dot plots

The second item in the post-study assessment that examined students' use of measures dot plots presented students with a series of three measures dot plots of the proportion of heads of a coin toss, but with the sample size not identified. To complete the task successfully students placed the three dot plots in order of increasing sample size based on a decreasing range of the distribution of measures, i.e., the correct sequence in Figure 4.27 was (i), (iii), and (ii). Methodology for the task is provided in Section 3.3.5.3, the item is attached as Appendix A.15, Q. 3, and the results are presented in Table 4.39.

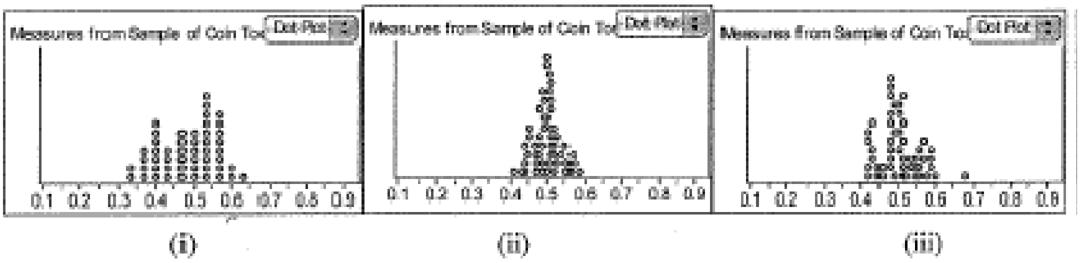


Figure 4.27. Mixed-up measures item post-study assessment Q. 3.

Table 4.39 shows that only 66.7% of male students and less than half of female students (45.7%) gave the correct, and relational, response of (i), (iii) and (ii). One quarter of the female students (25.7%), but only three male students (16.7%) reversed the sequence completely by incorrectly equating the largest distribution of measures with the largest sample size. This was considered a multistructural response because the student demonstrated an awareness of a relationship between the two, but had inverted the relationship. Three students confounded the sample size sequence entirely and placed the dot plots in no apparent logical sequence. Two male and two female students confounded the number of measures depicted on the graph (the students counted the number of measures present) with the sample size. Based on the data in Table 4.39 and the evidence presented in Section 4.5.3.4 students found the relationship between sample size and the distribution of the measures somewhat more difficult than the locating the centre of the distribution of the proportions of heads of a biased coin.

Table 4.39.
Students' Responses to the Mixed-up Measures

Response	Male students		Female students	
	No.	%	No.	%
No response	0	0.0 %	6	17.1 %
Incorrect unclassified	2	11.2 %	2	5.7 %
Confounds sample size	1	5.6 %	2	5.7 %
Reverses correct order	3	16.7 %	9	25.7 %
Three dot plots in correct order	12	66.7 %	16	45.7 %
Total	18	100.0 %	35	100.0 %

4.5.3.6 Students' post-study questionnaire items

The post-study questionnaire included examination of two general concepts of measures developed through the graphical representations of the proportion of heads. Item 1 "Increasing the sample size reduces the spread of the Measures" and Item 8 "The average of the measures is often the same as the expected value." Methodology for the questionnaire is presented in Section 3.4, and the results are presented in Table 4.40.

Students had little difficulty in responding to Item 1 and 94% of male and 91% of female students responded correctly, which was in contrast to the Mixed-up measures item (Table 4.38). Students found Item 8 more troublesome with only 17% of male and 57% of female students providing the preferred responses of agreeing or strongly agreeing that that the average of the measures is often the same as the expected value.

Table 4.40.

Post Student Questionnaire Sample Size Conceptual Items Proportion of Correct Responses

Post-study questionnaire item	Male students (n=18)		Female students (n=35)	
	No.	%	No.	%
1. Increasing the sample size reduces the spread of the Measures.	17	94%	32	91 %
8. The average of the measures is often the same as the expected value.	3	17%	21	57 %

4.5.3.7 Detailed study

Parts F and G were companion tasks. Part F (presented earlier in Section 4.4.16) varied the approach of the classroom study somewhat and examined students' interpretation of a graph of the more familiar cumulative proportion of heads plotted against the sample size. In Part G, presented here, students examined the proportion of heads, but with the data represented as a measures dot plot at sample sizes of 50 and 500 tosses of a coin. Students compared the measures dot plot with the cumulative proportion of heads representations and identified their preference as the most meaningful representation. Methodology is presented in Section 3.7.8, the worksheet is attached as Appendix D.1, Part G and an example of the dot plots is shown in Figure 4.28.

Students assembled the two Fathom measures dot plots of the proportion of heads at sample sizes of 50 and 500 tosses of a coin. In contrast with the classroom study Fathom was used both to generate and display the data as measures dot plots – in the classroom the dot plots were constructed manually. Students assembled the measures dot plots supervised closely by the researcher, because the objective was not to teach students to use Fathom features more advanced than those used in the classroom study.

Students' preferences for the graph format were divided evenly: approximately one third of students in the detailed study preferred the data represented as the cumulative proportion of heads, one third of students preferred the measures dot plot, and the remaining one third of students chose the format according to the circumstances (Appendix D.8).

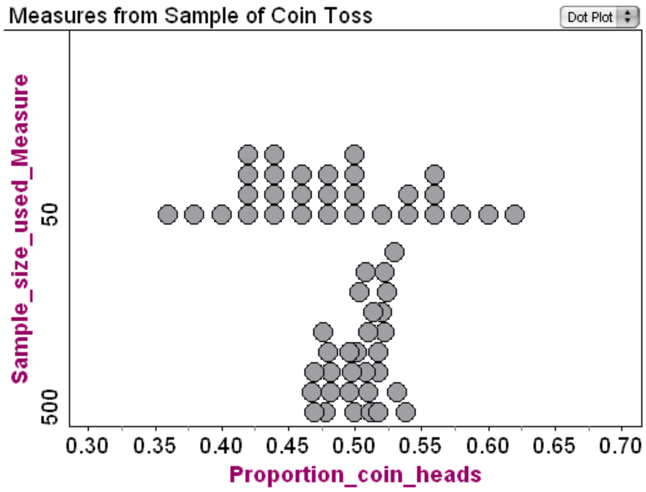


*Figure 4.28.*Illustrative example of the proportion of heads of 50 and 500 tosses of a coin generated and displayed using Fathom.

Students E1709S, N2701B and Y0304T preferred the cumulative proportion of heads representation because they thought the graph was less abstract and it was easier to make the connection to the underlying data. Student E1709S's comments were particularly insightful, and the student's concerns regarding the high level of abstraction were conceivably shared by other students.

> I like the previous one [the cumulative proportion of heads graph]…well when we did this in class [the measures dot plot] I thought it was really confusing because it is graphs inside graphs, it's like surveys inside surveys kind of thing and that makes it confusing when you see it like this, they aren't all single ones, each one of those dots represents lots of other dots so it's like …makes it really confusing … to see it because you can't…when you see it you can't think about what it actually is because there is a lot more behind it, so it is hard to see it all at once… [Student E1709S]
>
> The […] one, everything between, sort of thing and you couldn't get, you couldn't see it…[pause] it wasn't as easy to look at as this one. [Student Y0304T]

Students who preferred the measures dot plot, described here by students as the "50 and 500" in reference to the sample size used, appeared to recognise the underlying structure of the data distribution. The disadvantage of this approach appeared to be the higher level of abstraction, but it has the advantage of fewer features to distract and it may reveal the key features of the spread of the proportion of heads.

> Umm I think the 50 and 500 [Part G] …cause with the other graph, like I said from the one hundred to the three hundred, it was like really crazy, but with the 50 you can still see a shape but it isn't that changeable as much, and with the 500, it shows how even it actually is…if that makes sense. [Student S100J]
>
> This one [graph of the measures] shows the variation better, you can see where uhmm, the numbers [measures] are stacked up on each other so you can see how many there are, how many the same, at lot easier than the line graph (law large numbers) [Student Y1504L]
>
> This one [Part G, sample size 50 & 500] is holding heaps more information than the other one because…the other one [Part F, the cumulative proportion of heads] just represents just one [one run]. I think the other one would be easier to explain, but this one is a lot better for more information, like to show it all [Student R1610A]

Two students saw advantages in both graphs and that the choice would depend on the situation. This was a sophisticated and thoughtful response, because it indicated a versatile approach and choice of data representation based on what was most effective mathematically, rather than a choice based on what graphical representation about which the student felt most personally confident.

> Yeh it give you an idea, of the range yeh, because the range is easier to see and also the point where the mean is ... it all depends what you are looking for, whether it is biased or not for the line [expected value] ahrm…but if you want an idea of the range…sort of whether the extremes are and the range…the second one, with the dots. [Student T0612M]
>
> I prefer the 100 to 600 graph, just easier to read [more conventional format too] with the lines, but yeah, but [with the measures dot plot] it is easier to see how many there are a big group, with the line you can't tell there is a big group….[Student N2701B]
>
> Yeh, 500, whereas this one you compare directly the one hundred and the 500 in this case…ohh…between the 50 and the 500….hmm this one allows you to see, uhmm, …the way it went this way and that way, but the other one [Part F] allowed you to see it as you did it…[Student Y0304T]

4.5.3.8 Colleague teacher interview

The colleague teacher of the female class commented on the poster-sized dot plot of the fairness measures used in the investigation of the fairness of the dice. The teacher's observation indicates that the measures dot plot was well-received by the students, and that contributing their own data and consideration of their own data in relation to fellow-students helped create ownership and understanding of the data. Such observations, however, say as much about the pedagogical approach as the dot plot.

> They really liked it when you did the dot plots, putting their own data using stickers on the plots. They liked doing that because they were doing the graph. They had a physical spot, and they were looking at that in relation to everything. I thought that was really good. That was far more effective than just producing a graph, they actually made it. That's a good technique, and I would use that in future classes. They were really quite focussed on what they were doing and I thought it was a good way of doing it. [Colleague teacher of the female class]

4.5.3.9 Summary of findings for measures dot plots

The classroom study provided a learning trajectory from the first use of the dot plot to tasks given in the final assessment. Students had little difficulty in interpreting the dot plot graph in the pre-test female race-times data analysis task (Section 4.5.3.1), and this suggests that the dot plot format – as distinct from the measures dot plot – was comprehensible to most students.

The first assessment task interpreting measures dot plots was where students compared the fairness of the home-made, the factory-made and the Fathom die measures dot plots simultaneously (Section 4.5.3.2). Students, particularly the female students, produced sophisticated responses. This was a complex task; students were unlikely to have encountered a task interpreting three graphs simultaneously and of such complexity earlier in school. This was also a highly supported task with a clear practical objective of comparing the three dice.

The subsequent classroom activities were supported less explicitly by the researcher as the study sought to develop generalisable re-sampling skills in students, The effect of sample size on the fairness measure (Section 4.4.3) and the Fathom virtual 50 & 500 tosses of a coin (Section 4.4.6) were developmental activities that were designed to promote students' intuitive sense that the spread of measures decreased as the sample size increased, and the large population sample size model activity (Section 4.4.9) sought to extend these informal notions to quantifying formally the spread of measures. That the activities were only modestly successful in developing students' understanding of the concepts was shown by the subsequent homework item where more than half (57%) of the female students and one quarter (25%) of the male students misapplied the principle from the %fairness measure to the new context of the proportion of heads measures dot plot.

At the conclusion of the study it was clear that the two re-sampling principles that the spread of the measures decreases and the centre of the distribution of measures approaches the expected value as the sample size used to calculate the measure increases, were not robustly understood by students. Students' understanding was inconsistent and uneven – students applied the two principles successfully in the badly biased coin (Table 4.37) and the post-study questionnaire Item 1 (Table 4.42), but only half of the students completed the Mixed-up measures item successfully (Table 4.38). When considered from the perspective of the situated abstraction framework students demonstrated constructing and extracting, but were largely unable to demonstrate building-with to apply their knowledge to new tasks. Students' understanding of measures dot plots throughout the study was fragile and their knowledge developing. The detailed study examined students' preferences for the graph format most meaningful to them. One third of students preferred the more familiar cumulative proportion of heads graph, one third preferred the measures dot plot, and one third of students saw advantages in both graphs and said their choice would depend on the context. The students who preferred the measures dot plot and those whose choice would depend on the situation appeared to have developed some confidence in the use of measures dot plots. Different graph types promote different aspects of the data: that is their purpose. Compared with the cumulative proportion of heads graph, the measures dot plot eliminated much distracting material, but it also increased both the complexity (more than one series of a coin toss), and the level of abstraction (involving a calculation of a derived statistic). The measures dot plot was abstracted and extremely information dense.

4.5.4 Students' relationship with Fathom

The third and final aspect used to examine Research question 3 has, as a common theme, students' relationship with Fathom. This relationship is considered through: students' development of procedural use of Fathom in the classroom; four specific instances where Fathom appeared to have acted, through instrumentation, to promote students' learning; students' perceptions of Fathom taken from the detailed study interviews and four items from the post-study student questionnaire; students' recall of basic Fathom skills after a six week break such that might occur in a ordinary school year where the software is used intermittently; and the interviews with the colleague teachers.

4.5.4.1 Students' procedural use of Fathom in the classroom

Students were not expected to become proficient users of Fathom; they were provided with the opportunity to acquire only sufficient procedural skills to complete the tasks such that Fathom was not a constraint on learning. This was designed to allow students to focus on the underlying mathematical concepts, rather than to dedicate intellectual effort to learning the software or being distracted by the software. Evidence for students' development of procedural use of Fathom is provided by re-examination at four approximately equally spaced points in the teaching sequence at Lessons 1, 3, 5, and the post-study assessment (see Table 3.1).

The first activity, Lesson 1, was students' introductory use of the software. This was an exploratory data analysis task where students were asked to make and interpret a graph of the New York marathon race times data set (Section 4.2.5). This was not simulation, but the activity introduced the modules and key terminology common to Fathom data analysis and simulation. Students worked in pairs; one student in the pair was instructed first by the researcher and in turn this student assumed the role of tutor, explaining the use of the software to their tutee companion. Both the journal and the colleague teacher noted that students needed little support to use Fathom productively. Table 4.41 shows that 72% of the boys and 58% of the girls completed the task and provided submissions. The lower level of responses by the girls could be explained by a journal entry reporting that insufficient time was available to complete the task. The general observation in the class was that Fathom was being used productively mathematically within one lesson. Students' submissions are presented in Appendices F.2 and F.3

The second activity, Lesson 3, was students' first Fathom virtual die simulation and their second opportunity to use the software (Section 4.3.5). Students were provided with a Fathom die collection and a worksheet that included pictorial screen-grabs and instructions to complete assembly of the die simulation (Appendix A.6). The researcher first demonstrated how to assemble the simulation, and then the students used the worksheet to assemble a simulation. Table 4.41 shows that all of the male and 64% of female students provided submissions that showed that the simulation was assembled and used to complete all aspects of the task. The students had little difficulty in completing the tasks of assembling the die simulation correctly, using the simulation to collect data, and calculating manually the fairness measure for the Fathom die. With students' second use of Fathom the demonstration and detailed worksheet were sufficient to allow students to attend to the mathematical concepts, rather than the simulation. The students' submissions are presented in Appendices F.4 and F.5.

The third activity presented here, Lesson 5, was the activity "The effect of sample size on the fairness measure" (Section 4.4.3). Two worksheets were used in this activity: the male class used the standard worksheet format (Appendix A.8) of detailed step-wise instructions and diagrams of screen-grabs of a functioning simulation (e.g., Figure 3.1), and the female class used a simplified Fathom worksheet presented as a series of one line instructions (Appendix A.9). Table 4.41 indicates that 89% of the boys, but only 42% of the girls provided submissions that showed the simulation was assembled and all elements to complete task. The simplified Fathom worksheet was provided to the female students with the perception that students had acquired sufficient skills such that they no longer required detailed instructions. This perception was incorrect: the female students could not interpret the simplified worksheet correctly, and they assembled the simulation only with considerable individual support from the researcher. The female students became off-task and frustrated, and the underlying mathematical purpose of the lesson was ultimately lost. In contrast, the male students, using the detailed worksheet with diagrams assembled the simulation successfully. A worksheet using the simplified format may lie outside the limit at which students could use Fathom productively at that stage of their development. Students required the cues from the detailed worksheet to assemble the simulation correctly. In this instance Fathom, with instructions in the simplified form, was a constraint on learning, not an affordance. The class sets of students' responses are attached as Appendices F.6 and F.7.

The fourth activity was a post-study final assessment of students' basic Fathom skills. . The assessment was conducted under traditional examination conditions and students completed the task independently. To complete the task successfully students assembled a die simulation, changed the sample size from the default sample size of 10 to a sample size of 60, presented the data in a summary (i.e., a table), included a comments box, and answered three questions that required students to interpret the data. The task, although elementary, was designed to assess whether students could assemble and use a basic simulation independently. The methodology for the task is presented in Section 3.3.5.7, the task is attached as Appendix A.16, and students' responses are attached in Appendices F.8 and F.9. Table 4.41 shows that all boys and all but five girls provided a complete response. Several of the boys demonstrated their skills and extended the task to include the formula editor (Figure 4.4).or other sophisticated features.

Table 4.41.
Students' Development of Procedural Use of Fathom in the Classroom

Activity	Male students			Female students		
	No. submitted	No. simulations assembled correctly	% class simulations assembled (n=18*)	No. submitted	No. simulations assembled correctly	% class simulations assembled (n=33*)
New York marathon – introduction to Fathom	14	13	72 %	22	19	58 %
Fathom virtual die – first Fathom simulation	18	18	100 %	21	21	64 %
The effect of sample size on the fairness measure	17	16	89 %	19	14	42 %
Fathom basic skills test	18	18	100 %	31	26	79 %

* Nominal lesson attendance

Table 4.41 shows that the proportion of submissions by the female students was consistently and substantially lower than the proportion of submissions by the male students. This suggests that the software, the topic, or computer-based learning was valued less highly by the female students, or that female students were less confident than the boys about submitting their work for assessment.

The low level of submissions by the girls to the activity examining the effect of sample size on the fairness measure could be explained by the simplified worksheet used, by the students, at that stage of their use of Fathom the simplified worksheet did not provide enough information for them to assemble the simulation correctly.

The type and extent of support provided to students appeared to be the key factor influencing students' success. This support was reduced purposefully as the study progressed, and in the final assessment task students worked independently. The complexity of the tasks changed also, but an assessment of complexity of the four tasks is problematic other than the complexity increased modestly along with students' developing knowledge of the software.

The learning support and the social nature of learning in the classroom environment are recognised in instrumental genesis as orchestration. Three of the four examples of orchestration were successful, and one of the four was not successful (Table 4.41). The three effective activities were the introductory exploratory data analysis task, the Fathom virtual die simulation, and the post-study assessment. Orchestration using a tutor/tutee was effective. In all instances where the detailed worksheets were used orchestration was successful. By the conclusion of the classroom study and in the post-study assessment all of the male students and 79% of the female successfully assembled and interpreted a basic Fathom die simulation independently.

In the one instance where orchestration was manifestly not successful – the effect of sample size on the fairness measure activity in the female class – the support given to students was insufficient. From the instrumental genesis perspective this provides a point on the learning trajectory: at this stage of development students needed a worksheet with the cues of the graphics and the specific instructions to assemble the simulation successfully.

The post-study questionnaire invited students to comment on the worksheets. The standard worksheets were valued particularly by the girls with 80% either agreeing or strongly agreeing with Post-study questionnaire Item 16, "The worksheets with the diagrams were the most effective way to learn how to use Fathom." The boys were not as strongly positive that the worksheets were the most effective way to learn how to use Fathom – but only 22% disagreed the worksheets were not effective. The worksheets and students' approach to the software were described by one colleague teacher.

> Students seemed to cope well [with Fathom]. Some students religiously worked through the instruction sheets [worksheets]. Other students were so computer literate they were helping others and trying different things out [Colleague teacher of the female class]

4.5.4.2 Four aspects of learning promoted by Fathom

Instrumentation is the aspect of instrumental genesis that describes how the artefact acts upon the user to develop schemes and, in this study, describes how Fathom supported students' learning. Four examples are presented of how Fathom acted on students' thinking to support their transition from (a) a preference for frequency to proportional data representations that may indicate a shift to higher-level proportional thinking, (b) language use of “tossing a coin” to “sample size,” (c) small to large sample sizes, and (d) interpreting a graph to choosing a graph that “tells a story.”

Fathom appeared to promote students’ transition from frequency to proportional data analysis. The detailed study Part C examined students’ preference for numeric data representation and how that preference changed with sample size. Students were provided with a coin simulation and three summaries that presented simultaneously the data of a coin toss as a frequency of heads, a proportion of heads, and a percentage of heads. The sample size was changed progressively from the default sample size of 10 to 30, 74, and 273, and students expressed their preference for a data representation as “which one they looked at” for each of the four sample sizes. Appendix D.9, Part C shows that at the default sample size of ten eight of the twelve students (75%) in the detailed study preferred the data expressed as relative frequency of heads and tails, e.g., “5 heads and 5 tails.” At a sample size of 30 most students preferred the relative frequency, but students hesitated slightly as they calculated half of 30. At a sample size of 74 only four students (25%) expressed their preference for a relative frequency, At a sample size of 273 all twelve students had changed their preference to a decimal fraction e.g., “point 473.”: it was too difficult to calculate mentally half of the odd number 273. The shift in students’ preferences from frequency to proportion corresponds to a shift from additive to higher-order proportional thinking, and this shift occurred because of the large sample sizes available with virtual simulation.

Fathom appeared to promote change in students’ use of language from the informal tossing a coin to the mathematical formal expression of sample size. In the classroom study students used the informal expression “rolled the die 30 times” for the physical die or “tossed a coin” for the physical coin, and the expression “sample size” became common only when the Fathom simulation was used. In the detailed study students invariably used the shorthand expression of the numeric sample size of, for example, “300.” The terms were often used informally and imprecisely. Students did not describe a virtual coin simulation as “tossing a coin” (Appendix D.9). In a Fathom simulation there is no physical act of rolling a die or tossing a coin: the act is the more mathematically accurate “taking a sample.”

> I’ll go 5000 [when choosing a sample size] [Student R2408I]

> At 600 it's pretty much close to being completely fair. [Student S1001J]
>
> For the middle 300 it stabilises a bit. [Student R1610A]

Students' use of the expression sample size was observed indirectly as students completed the detailed study tasks. This suggests the expression was becoming part of the simulation that was supported by Fathom.

> I put the sample up to 50 and it is still swapping between the two [Student E2611G]
>
> I'm going to change my sample size [Student Y0706J]

Students used the expression sample size when discussing the concepts surrounding sample size. In discussing sampling concepts students' focus had shifted from the simple act of simulation and they were beginning to abstract the underlying concepts.

> It's a lot more efficient to go for a larger sample size [Student Y1504L]
>
> As the sample size grows the plus and minus gets smaller [Student N2701B]

Students did find the distinction between sample size used to calculate measures and the number of measures collected discussed earlier difficult, but they did appear to demonstrate development of use of the term sample size and this development appeared to be supported by the larger sample sizes and virtual simulation offered by Fathom. Such development may demonstrate a higher level of abstraction because of the disconnection between the act of simulation and sample size, and sample size is a general term that has wide application.

Fathom appeared to promote change in students' focus from small to large sample sizes. In Part D of the detailed study students were provided with a series of three, potentially biased Fathom coin simulations, and students were asked to determine whether the simulations modelled fair dice. The task also provided the researcher with the opportunity to examine both students' intuitive sense of sample size and their sample size strategies. The three dice, presented in sequence, were subtly biased with expected values between 0.48 and 0.53 proportion of heads, and the sample size was set initially to ten. Students made modest increments in the sample size of ten or twenty. Students were prepared to make a decision on whether the coin was biased using sample sizes within the range of 40 to 140 – sample sizes not dissimilar to a physical simulation. One student pair made a decision on bias at a sample size of 420 (Appendix D.9). When asked the sample size they would be content in assessing fairness two student responded:.

> Maybe 40…[Student Y1504L]
>
> I've done 150 [Student N2701B]

Drawing a formal robust statistically significant conclusion regarding bias of 3% required a sample size of approximately 1000. In Parts F and G of the detailed study students examined sample sizes of 50 – 600, and in Part H students explored sample sizes up to 5000. In these activities students' sense of sample size increased as the use of large sample sizes became commonplace. Students said initially:

> I'd say between 250 and 500. [Student N2701B]
>
> Between 500 and 1000. [Student Y1504L]

But as the activity progressed:

> Hmmm, I think I'll use 2000. [Student Y1504L]
>
> I have just gone up to 3000 and [….] the result is getting closer. [Student S1001J]

Fathom appeared to promote change from interpreting a graph to seeing a graph as a picture that tells a story. In Part E of the detailed study students were provided with a coin simulation and asked to construct a graph of their own choosing with the expressed purpose of creating a graph that "tells a story." With one exception students in the detailed study initially chose a simple frequency graph – students appeared to interpret the task as construct any graph. Students Y1504L, S1001J, and Y0706J provided examples of initial responses; the three students were displeased with their work.

> Mine is just the amount of times each face has come up. It doesn't give me any information on trends or anything like that. [Student Y1504L]
>
> I have done number of heads. I went to the proportion of heads first, then as it is a bit too confusing, so I think the number of heads is a bit better. [Student S1001J]
>
> I am just going to try the number of heads maybe ….oh, I like this one better….the frequency of toss, the number of heads and tails. [Students Y0706J]

Students were asked to make a second attempt. Several students' used trial-and-error until they obtained the desired appearance, or not, in this particular instance.

> Yea, ahrmm, first I put in times tossed and proportion of heads and then I thought…no…I then put in toss and it came up with two big bars, and I put in

> number of heads as well, sort of …and I don't get that now. [laughs] [Student S1001J]

Several students, possibly taking the cue from the classroom study, chose graphs of proportion of heads against the number of times tossed. Fathom largely eliminated the time required to construct a graph, which liberated students to focus on interpreting the graph, or interpreting a number of graphs in rapid succession as students sought to make sense of the data set. This was a complex task as students considered the data and the graph simultaneously, but a task facilitated by Fathom.

4.5.4.3 Detailed study – students' perceptions of Fathom

Students' perceptions of Fathom were explored informally in the detailed study. Common themes were that the software made it easier and quicker to complete tasks, such as constructing a graph or collecting data using a simulation. One student described the software as simple; this was not said as a criticism, and it is consistent with the research literature that advocates uncomplicated software that minimises potential distractions to learning.

> I thought it would be more advanced, more features to it [the software] if like, it was actually quite simple. [Student R1706D]
>
> It was easy to draw graphs and that, it wasn't really complicated or anything [....] table were easy to build up…just chuck it in [drag-and-drop] and it did it itself. [Student N2701B]

The software had features or aspects that students found difficult, but there was no apparent consistency or common issue troubling students. One student was particularly troubled by management of the Fathom workspace, and other students expressed difficulties recalling sequences of steps.

> I just remember moving things around a lot [on the computer], that's the main thing I remember. […] stuff piling up on each other really annoyed me [….] slowed it down a lot [Student N2701B]
>
> I occasionally got lost…with the dropping action, what would drop, how to get the attribute to be visible, how would I get a certain thing, say Face [an attribute] from the die one, onto the graph … I was not entirely sure and I had difficulty finding that…[Student N2610H]
>
> You sort of forget some keys you need to hold down sometime (laughter from other student)…exactly where you have to drag it … but it is all sort of easy like basically just little details you forget sometime. [Student S1001J]

Two female students found the software difficult to use on occasions, but their tone suggested a lack of fluency in using the software, rather than frustration in being able to use the software at all.

> I can't think of anything bad about the software, just learning how to use it was difficult. [Student E1709S]
>
> Yea getting up those little tables….and changing numbers and stuff…hard to remember sometimes. [Student R1610A]

The animation feature was a setting used when taking a sample or collecting measures that showed each individual sample or measure being taken. The feature was annoying to the students at large sample sizes because it slowed the simulation, but at large sample sizes it was not obvious that the animation feature was on.

> The animation on was annoying…couldn't tell any difference I couldn't tell if it was on or off…it didn't look any different [....] it just took ages…we didn't realise it was on…it took forever for it to come up. [Student R1610A]

4.5.4.4 Students' post-study questionnaire items

Four items on the post-study students' questionnaire examined their perceptions of Fathom (Table 4.42). Item 3 stated simply "Fathom was easy to use." Male students found the software easy to use with 88.9% of the male students agreeing or strongly agreeing, and no male student finding the software difficult to use. Female students found the software more difficult to use than the male students with slightly more than half of the females (52%) providing a positive response, 31.4% "maybe," and almost a quarter of female students (22.8%) finding Fathom difficult to use.

Students' responses to Item 29, "I would be willing to use Fathom again," paralleled students' perceived ease of use with 94% of male and 74.2% of females agreeing or strongly agreeing to be willing to use the software again. Only 5.7% of both male and female students were unwilling to do so.

Simulation software offers a mechanism to support the development of intuitions without emphasis on a formal mathematical approach. The role of simulation in the development of intuitions was explored through Item 4, "Even if I am not confident of the mathematics using a simulation can give me a 'feel' for the solution." Male students supported this statement strongly with 77.9% either agreeing or strongly agreeing. The item was less strongly supported by female students with less than half (45.7%) either agreeing or strongly agreeing; but a further 37.1% of females gave a neutral response suggesting that simulation may offer this opportunity.

In the two schools the spreadsheet MS-Excel was a readily available and potentially familiar alternative to Fathom. Item 18, “Fathom is too frustrating and I prefer Excel,” allowed students to state their preference: 94.5% of male students did not find Fathom frustrating and preferred Fathom over Excel. Female students were less positive, but still indicated a strong preference (68.6%) for Fathom, with 28.6% indicating no preference. This question presupposed exposure to Excel, but students’ pre-existing knowledge of Excel was not assessed.

In all four post-study questionnaire items that examined students’ attitude to Fathom the female students considered Fathom less favourably than the male students. Although the four questionnaire items specifically considered Fathom the apparent gender-based difference in attitude was consistent with students’ attitude to computer-based learning reported by Vale and Leder (2004), and it is also consistent with the lower level of submission of work samples by the female students in Section 4.5.4.1.

Table 4.42.

Post student Questionnaire Students’ Attitudes to Fathom

Questionnaire item		Disagree or Strongly disagree %	Maybe or neutral %	Agree or Strongly agree %
3. Fathom was easy to use	M (n=18)	0.0 %	11.1%	88.9%
	F (n=35)	22.8 %	31.4%	52.0 %
29. I would be willing to use Fathom again	M	5.7 %	0.0 %	94.0 %
	F	5.7 %	20.1%	74.2 %
4. Even if I am not confident of the mathematics using a simulation can give me a ‘feel’ for the solution	M	16.7 %	16.7 %	66.7 %
	F	17.1 %	34.3 %	48.6 %
18. Fathom is too frustrating and I prefer Excel	M	94.4 %	5.6 %	0.0 %
	F	68.6 %	28.6 %	2.9 %

4.5.4.5 Detailed study – students’ recall of Fathom after six weeks

The detailed study provided the opportunity to examine students’ recall and use of basic Fathom procedures after a six week period when the software was not used – a period that might occur with intermittent use in schools. The study also provided students with the opportunity to make any observations and discuss any difficulties they had when using Fathom.

Students were presented with a series of eight tasks, and within the eight tasks students were assessed on their ability to demonstrate four core skills essential for a Fathom re-sampling simulation: taking a sample, creating a summary (a table), changing the sample size, and creating a graph. Each procedure required a series of steps, but none of the operations required more than four steps to complete. Two of the operations, creating a summary and a graph, required an intelligent selection of the attribute to produce a meaningful data representation that would support analysis. Creating a summary and creating a graph required the same procedure, so if the student were able to perform one procedure, they were likely to perform the other.

Students attempted to perform independently or were given support if unable to do so. The students were assessed on a three-tiered scale of completed “by self” / “after demonstration” / “after demonstration and support.” If the students successfully demonstrated the core skill independently they had recalled the procedure from the classroom study, the procedure was sufficiently intuitive, or there was limited support from the researcher. In all instances students completed the tasks without significant difficulties that might cause delay in the classroom. Methodology for the tasks is provided within Section 3.7, the worksheet is attached as Appendix D.1, and the results are presented in Table 4.43.

Table 4.43 shows that students had little difficulty performing each of the four core skills after the skill had demonstrated. The most troublesome core skill was changing the sample size, which is a procedure that requires four separate steps. Students needed some support to be able to use the basic features of taking a sample, creating a summary and a graph, and changing the sample size. Data for these items were collected from two of the three pairs of male students only.

Table 4.43.
Detailed Study Students' Completion of Core Skills

		By self	After demonstration	After demonstration and support
Take a sample (Part A)	M (n=4)	0.0 %	100.0 %	100.0%
	F (n=6)	33.3 %	100.0 %	100.0 %
Create a summary /table (Part B)	M	25.0 %	100.0 %	100.0 %
	F	33.3 %	16.1%	100.0 %
Change sample size (Part C)	M	0.0 %	75.0 %	75.0 %
	F	0.0 %	50.0 %	66.7 %
Create a graph (Part E)	M	75.0 %	100.0 %	100.0 %
	F	83.3 %	100.0 %	100.0 %

4.5.4.6 Colleague teacher interviews

The two colleague teachers noted the ease at which students used Fathom, and neither identified any specific or common difficulty students had with Fathom. The use of the detailed worksheets appeared to support students' use of the software.

> Students seemed to cope well [with Fathom]. Some students religiously worked through the instruction sheets [worksheets]. Other students were so computer literate they were helping others and trying different things out [Colleague teacher of the female class]

Neither of the two colleague teachers had accepted the researcher's offer to develop familiarity with the software prior to the study, so their knowledge of the software differed little from the students participating in the study,

4.5.4.7 Summary of findings for students' relationship with Fathom

Students' development of procedural use of Fathom in the classroom was examined at Lessons 1, 3, 5 and the final post-study assessment of basic skills. Consistent with the study's position the objective of the study was to provide students only with sufficient skills to complete the tasks so that the software did not constrain learning. Fathom was essential for the study, but the software was essentially learnt incidentally to the study's activities. The study used a combination of one tutor-tutee activity and guided worksheets. The single instance where Fathom was not used effectively by a class was when a simplified worksheet was used by the female students, and this simplified worksheet did not provide enough information for the students to assemble the simulation. By the conclusion of the study students were able to assemble and run a Fathom coin simulation successfully.

Instrumentation was used to describe how the Fathom software artefact acted upon the user to promote learning. Fathom appeared to change students' preference from frequency to proportional data representations that may indicate a shift to proportional thinking; students' use of language from "tossing coin" to "sample size"; from small sample sizes of fewer than 100 to larger sample sizes of several thousand; and from students' interpreting a graph to choosing a "graph that told a story."

The male students' attitude to Fathom was more positive than the female students, but both groups largely found Fathom easy to use and were willing to use the software again. The gender difference in attitude to Fathom parallels that of male students more positive views of computer-based learning (Vale & Leder, 2004). The colleague teachers noted the ease of which students generally used the software.

In the detailed study conducted some six weeks after the conclusion of the classroom study – a time period that might occur in the academic year with intermittent use of Fathom – students had retained sufficient basic Fathom skills to complete tasks successfully.

4.5.5 Summary of findings for Research question 3

This sub-section considered whether Fathom re-sampling offers an effective learning opportunity for high school. Three aspects, peculiar to Fathom and re-sampling, of re-sampling terminology, Measures dot plots, and students' relationships with Fathom, were chosen for analysis.

The study examined students' use of the two key expressions thought essential in re-sampling of "sample size used to calculate a measure" and "number of measures collected." By the conclusion of the study students generally were unable to make a robust or consistent distinction between the two expressions. The instances where students made the distinction more successfully were the concrete tasks such as the introductory simulation that examined the fairness of the three dice and in the detailed study. The term measure was used to be consistent with the terminology used Fathom, but students' former use of the term was as a verb, i.e., to measure something rather than as the noun used here.

The study examined students' use of measures dot plots. Measures dot plots were used in this study to promote the two key re-sampling principles: that as the sample size used to calculate the measure increases the spread of measures decreases and that the centre of the distribution of measures approaches the expected value. Students had little difficulty analysing a dot plot of the pre-test, which suggested that the dot plot format was comprehensible to students. Students, particularly the girls, produced sophisticated analyses of the measures dot plots examining the fairness of the three dice, but students understanding demonstrated on subsequent less supported activities and independent assessment tasks were uneven and inconsistent. In the two instances of the fairness measure and the simulation of the public opinions survey in the detailed study the measures dot plot appeared to support students to identify the number of measures correctly.

The study examined students' relationship with Fathom. The guided worksheets developed for this study appeared effective in allowing students to assemble the simulation correctly, which in turn created the opportunity for students to focus on the underlying mathematical concepts. Fathom acted upon students and supported learning through instrumentation and four instances were identified. Within the study Fathom appeared to support students' transitions from frequency to proportional data analysis, promoted the use of the term sample size rather than "toss a coin," extend students' sense of sample size to numerically large sample sizes, and shift students' focus to a graph as picture that tells a story.

Several students reported some difficulties using the software, but the only consistent problem that occurred was that students initially found management of the workspace difficult. The male students favoured Fathom strongly, but Fathom was not as strongly favoured by the female students. Male and female students had little difficulty in using Fathom effectively within the first lesson, and students' apparent ready recall of the software in the detailed study after a six week period break suggested the software could be used productively in class intermittently.

4.6 Chapter summary

This chapter presented the data and provided a preliminary discussion of the information collected from the research items in the classroom and the detailed studies, the post-study questionnaire and the interviews with the two colleague teachers.

Research question 1 examined students' acceptance of the legitimacy of the Fathom simulation die using an investigation of a Fathom die simulation as a proxy. At the conclusion of the study the students had largely accepted the Fathom simulation as legitimate, and this study speculates that this acceptance eliminated one potential barrier to simulation, re-sampling, and to learning. This study provided students with the opportunity to participate in a process of statistical enquiry, to be introduced to re-sampling as a mathematical tool, and to develop and to use a formal statistic of the fairness measure. Students' demonstrated understanding of the fairness measure suggested that the mathematics involved was readily within the grasp of the students.

Research question 2 examined whether the large population sample size model was accessible to Year 9 students. A small proportion of students were able to apply the model to contextual tasks of public opinion surveys, but students generally found the mathematics involved challenging and students were not able to apply the model robustly. A follow-up test conducted approximately six weeks after the conclusion of the classroom study suggested that the large population model had not displaced students' preference for using an inappropriately large sample size, but fewer students favoured the sample size model of 10% of the population.

Research question 3 examined whether re-sampling offered an effective learning opportunity and mathematics tool for high school students. Three aspects of learning were chosen for more detailed analysis. Students' use of the specialised language of re-sampling terminology and of the distinction between sample size and the number of measures collected can be considered introductory only, and only approximately half of students were able to make the distinction successfully. Situated abstraction provided the framework to examine students' development and abstraction of mathematical information presented within the measures dot plots. Students had little difficulty in interpreting graphs presented as a dot plot of simple data sets, but they found independent and unsupported interpretation of measures dot plots more challenging. Female students who gave sophisticated analyses of the fairness measures of the three dice appeared to be supported by classroom activities conducted over several lessons, contributing and, owning the data, and access to notes taken during class discussion. The instrumentation aspect of instrumental genesis framework was used to examine how Fathom acted upon the students' thinking through to support learning. Fathom appeared to support students' transitions from frequency to proportional data analysis, from informal to more formal and abstracted mathematical language use of sample size, from small to large sample size, and to interpreting a graph as a picture that told a story. The study's use of detailed worksheets appeared to minimise Fathom as a potential impediment to students' learning that allowed students to attend to the mathematical concepts they investigated. The thesis now turns to a discussion of the results and a consideration of the implications of the study.

CHAPTER 5 Discussion and Implications

5.1 Introduction

This chapter examines the results and considers the implications of the study. The chapter begins with a consideration of whether the study was conducted as designed in the methodology (Section 5.2). The next three sections, Sections 5.3, 5.4, and 5.5, discuss the three research questions in sequence. Section 5.6 provides a general discussion of the study, Section 5.7 identifies the study's limitations, Section 5.8 considers the implications for teachers, researchers, software developers, and teaching resource developers, and the chapter concludes with a summary (Section 5.9).

5.2 Evidence that the study was conducted as intended

The overall framework for this study was provided by the six principles of the Scientific Research Approach (SRA), which are: posing significant research questions that may be investigated empirically, linking research to relevant theory, using methods to directly investigate the questions of interest, providing a coherent chain of reasoning, replicating and generalising across studies, and disclosing the research.

The three research questions are considered significant because the study examined elements of statistical literacy that are relatively unexplored and novel and which are topics not presently part of the current school curriculum. The research questions were investigated empirically using multi-methods of data collection of a classroom study that included student work samples, questionnaires, class discussion with students, interviews with the two colleague teachers, and a detailed study of six student pairs. A pedagogical approach recommended as best-practice by contemporary statistics education research guided both the choice of the concepts and the design of the classroom teaching sequence. Student work samples were assessed using the SOLO model. The study used methods that allowed the topic to be investigated directly using an approach chosen purposefully to be naturalistic to determine students' responses and development within the familiar classroom learning environment. The study provided a coherent chain of reasoning of the three research questions that examined students' beliefs, knowledge, and naive notions of the concepts through pre-testing; offered a program of statistical enquiry where the fairness of the simulation was examined objectively and considered students' development of the explicit determination of sample size; and concluded with post-study tests and questionnaire items that established students' post-study beliefs. Students' development of understanding of key elements of terminology and measures dot plots were examined using situated abstraction, and students' development of use of the software Fathom was observed using principles based on instrumental genesis. The scope to replicate and generalise across studies was limited by the resources available. Data were collected in two mathematics classroom environments in two high schools, but this was principally to provide responses from both male and female students, and should therefore be considered a parallel, rather than a duplicate study. The research is disclosed (published) here in this thesis, and in conference proceedings (e.g., Bill, Henderson, & Penman, 2010).

The study was conducted ethically under the Human Research Ethics Committee (Tasmania) Network (HERCS) H009790 and the Department of Education Tasmania (Ref. 672670). The approval for the final and closing HERCS ethics reports and the approval from the Department of Education Tasmania are attached (Appendices B.16 , & B.2). A scanned set of signed copies of consent forms from the parents and students participating in the detailed study are attached in Appendix B.14. Students in the classroom study participated by benign consent and did not provide signed consent. To the researcher's and the supervisors' knowledge no concerns regarding the ethical nature of the study or the personal conduct of the researcher were raised by the schools, the principals, the colleague teachers, parents or the students, HERCS or the Department of Education Tasmania.

Students engaged with the big statistical ideas of samples and sampling. Statistical intuitions, conceptual understanding, and a culture of enquiry and statistical process were cultivated purposefully through activities that examined students' beliefs of the simulator as fair and a frequentist approach was used to demonstrate the validity of the large population sample size model. Fathom, the technology in the study, was used in approximately 40% of class time. Formal mathematical experiences of substance were provided through the development and use of the fairness measure statistic, consideration of measurement and measurement error, an examination of graphical data representations, and the use of the large population sample size model. Whole-class classroom discussion played a prominent role in the classroom study, and one colleague teacher noted the importance of classroom discussion in this study.

> The best way of reinforcing your understanding is to explain it to someone else, and they can listen to others and it reinforces what they do know. I liked the other issues that were dragged out along the way. [Colleague teacher of male class]

Validity refers to the accuracy and credibility of the research findings, and whether the study investigated what it set out to investigate. This process was supported by multi-methods of data collection, triangulation, and prolonged time in the research environment of the classroom and the detailed study. This was a naturalistic study conducted in a normal classroom environment.

5.3 Research Q. 1. Developing acceptance of the Fathom simulation

Research question 1: How effective is a statistics education research best-practice based approach of scientific enquiry in developing high school students' acceptance of the Fathom virtual simulator?

The research literature (e.g., Batanero & Diaz, 2005) indicated that students were likely to bring to the study probability beliefs and misconceptions that could affect learning. Drawing on this literature the researcher speculated that a lack of acceptance of the Fathom simulator might potentially have presented a barrier to learning the mathematical concepts. The study sought to foster students' acceptance that the Fathom die was fair enough for the purposes of the students' mathematics class, and this was used as a proxy for students' acceptance of Fathom simulations more generally.

Did students accept the Fathom simulator? In short, the answer was largely "yes." In response to the post-study questionnaire Item 7, which asked whether Fathom die was "as random as it needed to be for the work we were doing," only one student of the 53 students disagreed. Many students (42.9% of male and 9.4% of female students) thought that the Fathom simulation was fairer than the factory-made die and argued that the computer eliminated the biases introduced by the way the die was rolled or the physical environment. Both colleague teachers felt the students had been convinced that the simulation was fair.

Scepticism and doubt regarding the die simulator persisted amongst a small number of students. This scepticism related to either simulation results that students found peculiar or more entrenched scepticism of electronic technology generally. Students' scepticism of the technology was expressed as a lack of transparency in the operation of the simulation, even though students had assembled, tested, and used a number of Fathom simulations. More generally the exploration of unintuitive results of probability simulations provides a legitimate reason for using virtual simulation, but persuading students to accept these simulation outcomes without having first developed students' acceptance of the simulation process may be problematic.

Students' confidence in the Fathom die relative to their confidence in a physical die was more ambiguous. Although students had accepted the Fathom virtual die, many still preferred the physical die. Approximately one third of both male and female students retained greater confidence in the physical die, one third had greater confidence in the Fathom die, and one third had no preference (Section 4.3.6). If students expressed a preference for the physical die, that preference disappeared when large sample sizes made use of a physical die impracticable. Students in the detailed study recognised the speed and efficiency of the Fathom simulation, but they noted the disadvantages that it required a computer and software, and the simulation needed to be assembled, but that physical die and coins are readily available. This could be interpreted as students choosing intelligently the most appropriate technology for the task at hand.

It could be argued that it was unnecessary for the researcher to devote precious class time attending to students' beliefs of the simulation if students had little doubt initially or were likely to accept the simulation readily. The teaching sequence, however, served a dual purpose simultaneously: developing acceptance of the simulation and providing opportunities for learning through statistical enquiry and purposeful mathematical activity. These were valuable learning experiences that provided an opportunity to develop skills that would find application elsewhere and provide a foundation for study in more senior years. The intent of this phase of the study could legitimately be re-cast as statistical enquiry, with only one of the outcomes being development of acceptance of the simulation.

Central to the instrumental genesis framework (e.g., Drijvers & Trouche, 2008) was the concept of schemes that were the mental processes needed to use the software tool for the task at hand. Acceptance of the simulation as "fair" was an example of such a scheme. A second example of a scheme was that students had begun to internalise the basic procedural use of Fathom, and students were developing the procedural knowledge to use Fathom effectively. The two schemes of beliefs and procedural use of Fathom have in common that neither proved a significant barrier to learning. The students, particularly the males, thought assembling the simulation was important in their acceptance of the simulation.

This study took a more guided approach to students' acceptance of the simulation than that of Pratt and Noss (2002) who examined the approaches students chose when using virtual probability systems. Pratt and Noss described virtual simulations as cueing unpredictability and unsteerability. In this study no student was observed using unpredictability as a test of the random behaviour of the simulation, which was consistent with Ireland and Watson (2009) who thought that it was the larger sample sizes of virtual simulations that encouraged students to look beyond short-run behaviour. The small proportion of students who remained sceptical of the simulation's randomness expressed concerns that could be classified as unsteerability, i.e., that the simulation was controlled by the computer or purposefully biased by the researcher or somehow artificially contrived. In observing students' use of resources to support their development of understanding Pratt and Noss also identified the development and use of what they termed a distribution resource (p. 471), which is formal recognition of the distribution inherent within the simulation. In this study this distribution resource was cued prior to the study (in the pre-test Q. 4) and students generally expected each face of a die to occur with a similar frequency. This item presented students with a data summary; the short-run sequence of the die roll was not available to them, so unpredictability as a test of random behaviour was not cued.

Students' development of belief in the random behaviour of a die can be traced through their responses to the initial, developmental, and final assessment tasks. Students' initial attitude to a conventional factory-made die defaulted to the attitude that the die was fair, and the results were attributed to chance. In the pre-test three students only were prepared to entertain the idea that the die may not be fair, and all three proposed investigating the die's fairness by increasing the sample size (Section 4.2.2). Students' unsophisticated responses could be explained partially by the few minutes available in the pre-test to respond. In the developmental activity using their (unfair) home-made die students examined the physical character of the die and the random behaviour of the die, and proposed criteria or methods to assess the die's fairness. Students had additional time, and stimuli from the activities and the worksheet questions that formally and mathematically examined the die's fairness, to provide richer responses. Several students proposed a mathematical model similar to the fairness measure used in the study (Section 4.3.3). On the final assessment task a large proportion of students, particularly the females, were arguing from the evidence in a sophisticated response using the fairness measure and three data distributions (Section 4.3.6). The high proportion of sophisticated responses provided by students was in contrast to studies reported by Watson and Moritz (2003) where the two researchers attributed students' low level of responses to classroom cultures where students did not test their beliefs of random behaviour in empirical trials.

The education research literature advocates introducing familiar physical simulations before virtual simulation (e.g., Watson, 2006), but much of that research was conducted with students somewhat younger than the cohort studied here. The time allocated to this aspect in a unit of study may be dependent on the student group. The colleague teacher of the all-girl class thought that the physical simulations could be shortened or even eliminated entirely. This was consistent with the researcher's journal that recorded that the female students' behaviour with the second physical activity using coins was largely of amusement, as the task had been done in more junior years. It was important not to insult students with activities seen as puerile, and this may have undermined the credibility of the activity. One physical activity with the die may have been sufficient. In contrast the male students may have benefited from the greater physical activity that the physical stimulation provided.

This study chose to examine the fairness of the Fathom die through a formal process of statistical enquiry, but the fairness could have been explored in a number of ways. A simple and obvious strategy would have been to compare the relative frequency with which each face occurred, but this strategy employs additive thinking only and would do little to promote mathematical meaning at the Year 9 level. Students in Year 9 extended mathematics – as shown by the study – were capable of a more sophisticated and mathematically rewarding approach. The study sought to promote a shift to higher-order thinking – from additive and multiplicative to distributional thinking (e.g., Garfield, Del Mas, & Chance, 2007). Comparing the relative proportions of a simulation requires multiplicative or proportional reasoning, but compares single numeric values only. Comparing the relative proportions through pooling and displaying the classes' results on a graph was designed to promote higher level distributional thinking. The most challenging task for students was comparing the three data distributions (Section 4.3.6). Prior to the study students' interpretations of distributions were possibly limited to interpretation of single distributions, rather than a comparison of data distributions. Students appeared to have benefited from the foundation activities using the GICS framework to examine the single distribution, of the data of a home-made die, and finally using GICS to compare distributions of the data from all three dice. Students' ability to communicate their understandings of distribution was critical to most of the aspects of distribution used to evaluate students' understandings of distribution.

This study examined Fathom simulation as a legitimate and effective teaching tool for high school, and part of the software's legitimacy lies with students' acceptance of the Fathom simulation as fair. Attending to students' acceptance of simulation may need to be one of the objectives in all activities (Watson & Moritz, 2003), because students' acceptance may be specific to a particular situation; that is, students' acceptance of the Fathom die as fair does not imply that the Fathom coin simulation used subsequently in the exploration of sample size was considered fair. This acceptance has at least two aspects. The first is addressing students' acceptance at the subjective level, and the second is an objective analysis that the simulation behaved as intended. Attending to acceptance at the subjective level may not be necessary, but acceptance at the objective level is an integral part of sound statistical enquiry and potentially a mathematically profitable classroom activity.

The study's finding in response to the first research question is that the students, using a process of statistical enquiry examining the fairness of dice, accepted the Fathom die as fair. This is supported both by student feedback and by the learning outcomes. This acceptance was not universal: some students' acceptance was qualified and students presented with a mosaic of beliefs. Students were highly engaged by the introductory activity examining the fairness of the home-made die. Testing the factory-made and the Fathom die re-enforced both the test procedure and the process of statistical enquiry, and the development and application of the fairness measure provided the students with a formal mathematical analytical framework. Arguably students may have accepted the simulation as legitimate without a process of formal investigation. For these students this study provided an opportunity to see modelled and to participate in a process of statistical enquiry, and statistical enquiry will find many applications elsewhere. For a small number of students there may be a link between students' assessed abilities with mathematics and their acceptance of the simulator, with less mathematically able students less likely to accept the simulation as fair. The possible link between lower levels of mathematical ability and acceptance of the simulation is a legitimate topic for further research.

5.4 Research question 2: An accessible sample size model

Research question 2: In what ways does the sample size model e = ± 1/√n provide an accessible method for high school students to explicitly determine sample size when sampling from large and infinite populations?

The study took from contemporary best practice statistics education the importance of cultivating accurate intuitions, conceptual understanding, and beliefs, and the criticism that earlier procedural approaches to statistics education may confound learning. Building on this literature the implications of this research question lay beyond simple procedural application of the model to include students' development of conceptual understanding of sample size and survey accuracy, and the application to contextual tasks. Students' use of the sample size model and their development of intuitive understanding were examined from the perspective of students' change in beliefs of sample size, use of models of a single statistic, use of the large population sample size model of $e = \pm 1/\sqrt{n}$, interpretation of survey accuracy, and application of the model to contextual tasks.

In response to whether the sample size model was accessible, the research study finding was inconclusive. The substantial development of understanding that occurred during the classroom study, particularly amongst the female students, was not sustained to the follow-up test conducted two months after the classroom study concluded. In the follow-up test students' understanding was only subtly more developed than the beliefs held prior to the study, and students did not incorporate significant development of understanding of sample size into their thinking permanently. The students who had used the sample size model correctly and gave sophisticated responses during the classroom study and on the post-study assessment may have applied the model largely from memory. In the follow-up test only two students mentioned the large population sample size model, so the objective of providing students with a convenient and more mathematically robust sample size model than the 10% of the population model was unsuccessful. Year 9 may simply be too early in students' academic lives to introduce the formal sample size model, or the classroom time studying too short for students to develop a robust understanding of the concepts. The entanglement of beliefs of probability noted in the education research literature and demonstrated in the study by students persisted (Batanero & Sanchez, 2005). Such responses highlight the importance of providing students with comprehensive sample size strategies for both small and large populations that they will encounter in ordinary life.

Students' development of beliefs about sample size when sampling from large populations was established through the national and state election item given first on the pre-test and on the follow-up test administered two months after the conclusion of the classroom study. Students' naïve understanding of sample size in a large population showed that students favoured a sample size far larger than used conventionally and a preference for using an even larger sample size when sampling from a larger population. The most favoured strategy on the pre-test, preferred by 62% of male and 55% of female students, was a sample size of 10% of the population. This is clearly impracticable for populations for national and state opinion surveys. Naïve sample size strategies also favoured a sample as a proportion of the population and, for the male students at least, "as large a sample as possible to improve accuracy." If an even larger (and impracticable) sample size of 150,000 had been offered as one of the multi-choice alternatives it is conceivable students may have chosen that alternative. On the follow-up test the most noticeable development was students' shift away from the alternative given of 10% of the population strategy to a sample size of 15,000, a change that could be explained largely by the impracticality of the 10% of the population strategy with large populations. Students gave very low level responses on the follow-up test, and it will be these beliefs and knowledge of sample size demonstrated in the follow-up test that students will likely take into ordinary life outside the classroom.

Students were unfamiliar with the relatively large populations, e.g., Hobart 200,000, and the large sample sizes used in the study; they seemed unaccustomed to sample sizes beyond fifty, and the sample sizes used may have been larger than the populations they had encountered previously. Students were familiar with die and coin simulations, but it was unlikely that students recognised these populations as infinite populations prior to the study. Virtual simulations provided the opportunity for students to examine sample sizes far in excess of that previously available to them using physical simulation. In the detailed study students were offered simulations set to small sample sizes, and their intuitive increments in sample size were far smaller than the sample sizes of thousands needed to complete the task.

The study included an examination of whether or not students were able to abstract mathematical meaning from the models used in this study. Students were able to abstract mathematical meaning from the single statistic of the fairness measure, and this suggested the fairness measure was accessible to them. Several students proposed statistics to measure a die's fairness similar to the model used ultimately in the study. The fairness measure homework items (Section 4.3.4), sought to deepen students' understanding by examining the mathematics within the statistic beyond simple application to examining the fairness of the dice, and this item did not trouble the many students who provided responses. Within the study of the fairness measure students exhibited the nested relationship of all three elements of situated abstraction that used their existing mathematical knowledge and understanding of a die's behaviour to construct, recognize, and finally building-with (Hershkowitz et al., 2001) to apply that knowledge to the fairness measure homework. Students had few difficulties with the, albeit simpler, mathematics associated with the fairness measure, and the fairness measure statistic appeared to be accessible to students.

The activity "The effect of sample size on the fairness measures" size (Section 4.4.3) was designed to provide a transitional activity between the fairness measure activities using dice and the subsequent formal exploration of sample size. The activity was not universally effective, and it confounded many students. The activity sought to promote the key concept that the spread of measures decreased with an increase in sample size, but also for some students promoted the erroneous understanding that the centre of the distribution appeared to shift to the left – as students submitted in the coin 50 and 500 tosses of a coin homework item (Section 4.4.5). The item illustrated that few students could, at that stage, abstract sufficient knowledge from the task to build-with and apply their knowledge in a different context. The emphasis given to providing a transitional activity between the die fairness measure activities and the subsequent sample size activities involving coin simulations may have been unnecessary or some alternative should have been used. In hindsight the study may have been more productive if the examination of the effect of sample size on the fairness measure had been postponed until after examining the effect of sample size on the simpler coin system.

The development of students' intuitive understanding of sample size needed greater emphasis than was given in this study. Earlier studies with students younger than those participating in this study focussed on the creation of sample spaces and intuitions of sample size, and several of these studies demonstrated development of understanding amongst students (e.g., Baker & Gravemeijer, 2004; Konold, Harradine, & Kazak, 2007). This study had a broader focus and sought both to cultivate intuitions and to develop formal notions of sample size simultaneously. Developing intuitions was an essential, but not a sufficient, learning experience because Year 9 advanced mathematics students should also be supported towards the formal mathematical approach that they are likely to encounter in more senior school years, in combination with sense-making and the development of intuitions. The importance of attending to students' subjective beliefs of probability and their intuitive sense of sample size was demonstrated by the simultaneous persistence of contradictory notions of using a sample size of 10% of the population for a large finite population and a sample size of typically fewer than 100 for the infinite population of a coin or die simulation.

The study examined whether or not students could abstract mathematical meaning from the sample size model. This provided students with their first exposure to functions of the form $y = 1/\sqrt{x}$. On the pre-test (Section 4.42) three quarters of both male and female students were able to substitute a small integer into a function of the form $y = 1/\sqrt{x},$ but the remaining students either confounded the two operations of square root and reciprocal, ignored the reciprocal operation, or did not provide a response. This suggests that some students' background knowledge of surds and reciprocal operations was modest. In the first classroom activity and one designed to demonstrate the utility of the model students had little difficulty calculating the margin of error when they first examined the large population sample size model (Section 4.4.9). Students' knowledge of the sample model was more fragmented than the fairness measure. While many provided evidence of constructing and recognizing the mathematical concepts within the sample model, few students were able apply the model to new tasks independently or to novel applications that lay beyond the tasks encountered in the classroom. For example, in the final assessments items of the mixed-up measures (Section 4.5.3.5) and the mathematics of the model (Section 4.4.11) students were largely unable to recall and use the model to determine the spread of measures or calculate the sample size for a given margin of error. When viewed from the framework of situated abstraction students were able to construct and recognise the mathematics within the model and apply the model in a supported environment, but only a small number of students were able to build-with and apply the knowledge independently to unfamiliar tasks.

Colleague teachers' opinions were divided on whether the mathematical model should be introduced and examined before the contextual sampling task, or the contextual task should be given first to demonstrate the need for a sample size model. The study's speculation that students would not arrive at the model independently was justified: few students were able to propose a mathematical representation for the simpler task of a fairness measure (Section 4.3.3), and it was unrealistic to expect students to develop the more complex sample size model independently. If the model were to be used, purposeful, or at least guided, introduction of the model was essential.

The study sought to cultivate students' intuitive sense of the accuracy of survey measurement through consideration of quantifying the variation associated with random variation. In the task Compare intuitive sense of a 50 tosses of a coin with a Fathom coin toss (Section 4.4.7) a student described the prediction of "19" with an observed "25" as "pretty close" when a simple, but formally statistically incorrect, calculation accessible to Year 9 students would consider the difference as an error of 24%. The desire to improve accuracy did not consider a meaningful or appropriate level of accuracy, or more simply "how accurate do I need to be?" In the follow-up test conducted two months after the study students' thinking had shifted from a strategy of 10% of the population to a strategy perhaps best described as the next practicable smaller sample size, but this shift was driven by the cost and practicalities of conducting a survey, rather than any formal consideration of meaningful accuracy. A proportion of both male and female students considered accuracy, but informally only, with the intention of maximising accuracy as opposed to choosing the accuracy appropriate for a particular situation.

The detailed study suggested that consideration of measurement accuracy in the context of familiar physical measurement of length, mass, or time provides a foundation for consideration of accuracy when sampling. Students bring to the classroom some intuitive sense of meaningful measurement that could be supported and developed formally in the mathematics classroom (and across the curriculum in quantitative subjects such as science, physical education and the manual arts). This sense could be extended to consider the accuracy as a proportion of a measurement, which would also provide an opportunity and justification for calculations involving percentages. The study’s use of notation (+/-) to describe accuracy may have been unfamiliar to students initially: on the pre-test a quarter (23.8%) of male and a third (30.3%) of female students did not provide a response to a multiple choice question that used the notation.

The study did not attend sufficiently to the accuracy of measurement. The worksheets, used to both teach and assess students, may not have dealt adequately with the concept of accuracy, or indeed the concepts around sample size in general. The accuracy of measurement, meaningful measurement, and the practical significance of measurement all have a place as part of sense-making and the interpretation of data. The accuracy of familiar physical measurement seems an essential prerequisite to consider the accuracy of sampling. Students were not appropriately critical of their own data, and this may reflect their lack of skill in considering measurement error.

One of the criteria of whether the model was accessible was whether students could use the model in contextual tasks. The Mt. Wellington cable-car item provided information on students’ development of understanding within the time-frame of the classroom study, which is the time-frame within which school assessment occurs normally. When presented first as a homework item (Section 4.3.4) students had, at that stage of the classroom study, exposure to large sample sizes, but not to the large population sample size model. The item differed somewhat from the national and state survey item because students argued from a position of justifying a given sample size of 900, rather than independently choosing a sample size. A task that nominated a sample size may have provided a cue that this was an appropriate sample size. Students’ naive strategies were predominantly that of a sample size as a proportion of a population. When presented again as a final assessment task many students produced sophisticated responses that included consideration of sample size, measurement error, and the consequence of measurement error (i.e., whether the outcome of the survey changes). Within the time-frame of the classroom study – up to the post-study assessment and the period assessed in schools normally – development of understanding had occurred (Section 4.4.14). Students’ development is consistent with, but not as extensive as, that reported by Smith (2004) who examined older college level students’ consideration of sample size. The Smith study reported considerable development of understanding of sample size concepts that included students’ replacement of the 10% of the population model with more sophisticated understanding of sample size.

The study sought to provide students with mathematical experiences of substance, and this study took the position that mathematical experiences in Year 9 should include mathematical modelling, use of a reciprocal function, and virtual simulation. The study included a number of activities of calculations involving surds and inter-conversion of decimals and fractions. The large sample sizes made feasible by virtual simulation simplified calculations and allowed use of numbers that had integer solutions and allowed simple inversion. For example, the square root of a sample size of 1600 is 40, and the reciprocal of 40 is 0.025. The most challenging aspects for students – 60% of female students did not attempt the task – was the algebraic manipulation required to determine the sample size for a spread of measures displayed on a measures dot plot. The mathematics within the sample size model was not explored extensively, and a failure to make connections with formal mathematics may have contributed to students' lack of development of understanding.

Fathom had two limitations that directly affected the exploration of sample size, which are discussed here rather than in the limitations section. The first limitation was that the sample sizes of 15,000 and 150,000 used in the national and state election survey item could not be modelled directly in Fathom: Fathom was limited to a sample size of 5000, and to generate a larger sample size the sample had to be accumulated batch-wise. A sample size of 1500 may be compared with the maximum sample size of 5000, but the margin of error at these two sample sizes are ± 2.6% and ± 1.41% respectively, and students may consider the difference as too subtle to have any practical significance. The second limitation was that Fathom was unable to support very large populations directly, i.e., where each individual is actually present in the data set. For example, a large population could be represented as a data set of 50,000 individual data points, but the random sampler operated too slowly to be useful in the classroom. Very large populations were modelled indirectly by random simulation that generated values, rather than sampling randomly from an existing population data set.

Given that the model was introduced with modest success, the question arises as to how the model might be introduced more effectively in high schools. The study took the principles from statistics education research of using technology in a new way to exploit the potential of the software tool (Ben-Zvi, 2001) so the study introduced and used re-sampling. This may not have been the most effective strategy. The value of the sample size model for Year 9 students may be as an extension of the Law of Large

Numbers activities, the exploration of a function of the form $y = 1/\sqrt{x}$, the practical application of surds, and an opportunity to practise basic number skills. Traditionally the Law of Large Numbers is used to demonstrate the principle that as the sample size is increased the observed value approaches the expected or population value, and studies such as Pratt and Noss (2002) showed that this principle was accessible to students much younger than the cohort in the study presented here. This suggests that students participating in this study may have benefited from the type of more advanced work offered by this study, so this study sought to extend the concepts to mathematising the Law of Large Numbers and quantifying the approach to the expected value. Virtual simulation allows the exploration of far larger sample sizes than is practicable with physical simulation, and this provides the opportunity to explore the sample size model used in the study at a wide range of sample sizes.

The model has clear application to quantifying the variation associated with sampling from the infinite populations of coin and die simulations, but the model's application to contextual tasks of sampling from large finite populations raises many additional conceptual issues that need to be addressed separately in the classroom. Examples of conceptual issues associated with sampling from large populations in this study were students' persistent beliefs that a sample size must be related to the population size and their perception of whether or not a computer simulation of an infinite population can model a large population contextual task effectively. One teaching strategy would be to use the sample model for infinite populations in Year 9 and provide a theoretical foundation for students, and then apply the model to contextual large population tasks in subsequent school years. A second teaching strategy would be to use the model as a bridging and companion task to sampling from a small population where the sample is a significant proportion of population. Small populations are conceptually simpler, and this may also support students because the populations used, a school for example, are contextual. The small population sample size model is, however, mathematically complex and Year 9 students may find simply substituting values into the model difficult. Small population sample models incorporate the widely-held – and correct – intuition that a sample as a proportion of the population is significant; it is only in very large populations that the influence of the sample as a proportion diminishes and can be ignored.

The study made only the first tentative exploratory steps towards use of the sample size model in high school. The model cannot be learnt in isolation from the cultivation of students' intuitive sense of the effect of sample size, and students at Year 9 may have the potential to accompany the development of intuitions with the formal mathematical exploration of sample size.

5.5 Research question 3: Fathom re-sampling as a tool for high school

Research question 3: In what ways does this study's pedagogical approach of using Fathom virtual simulation and re-sampling offer an effective learning opportunity for high school students? What affordances and constraints do students encounter?

Three aspects – peculiar to Fathom and re-sampling – of the key terminology of "sample size used to calculate a measure" and "the number of measures collected," measures dot plots, and students' relationship with Fathom were used to determine whether Fathom re-sampling offers an effective learning opportunity for high school students. These three aspects were intended to be an indicative, but not comprehensive, means of determining how Fathom should be introduced and used, as well as to identify the affordances and constraints that students encountered.

5.5.1 Re-sampling terminology

The statistics education research literature notes the importance of classroom discourse to promote learning and statistical thinking, and to support enculturation into the statistical process (e.g., Ben Zvi, 2004a). An essential supporting aspect is the vocabulary of statistical analysis (Baker & Gravemeijer, 2004), or conversely the lack of an appropriate vocabulary may be a constraint on learning. This sub-section examines students' use of two key re-sampling terms: "the sample size used to calculate a measure" and "the number of measures collected."

This study was students' first formal experience of re-sampling. Prior to the study students' experiences would have been to calculate one measure, such as a mean, from a sample size of *n*. Re-sampling introduced multiple measures, so students needed to be able to distinguish between the sample size and the number of measures collected. To make the distinction clear the two expressions "sample size used to calculate a measure" and "the number of measures collected" were used. What students knew formerly as sample size was in the context of re-sampling the more complex expression "sample size used to calculate a measure," and where students formerly had one measure only in this study they collected and recorded a number of measures.

Words that have a natural language meaning may have a more precise or different definition in a mathematical context. Students faced several transitions in the use of the terminology as they negotiated meaning from natural language equivalents to the formal use in re-sampling in the language-game (Meyer, 2009) of the mathematics classroom. The term statistic is used universally and was more familiar to students, but the term measure was introduced to be consistent with Fathom. In natural language the term measure was known to students as a verb, not as the noun used in the study. A statistic has the fundamental principle, which was defined earlier in this study, as "a number that represents a more complex set of numbers" (Section 3.2.9). The fairness measure was a statistic that was developed for a specific purpose but had limited application outside of the study, and the measure of the proportion of heads was familiar to students, but it may not have been also known as a statistic.

At the conclusion of the series of activities that examined the fairness of the three dice, students' understanding of the number of measures collected seemed clear. Factors that may have contributed to students' ability to abstract mathematical meaning included a clear sequence of the steps of roll a dice → calculate a fairness measure → contribute one measure to the class data set. There was a three-fold repetition of the task with the three dice tested in sequence, and there was a direct visible connection between data collection and analysis, where the class set was displayed prominently as a wall poster for an extended period.

In this series of activities individual students did not actually conduct multiple re-sampling: it was the class that re-sampled, not individual students. Individual students calculated one fairness measure only, which they then contributed to the class data set. The level of abstraction was modest: a clear connection existed between the data and the measure and the principle was not generalised. Students were, however, participating in, observing, and seeing modelled the re-sampling process – an essential and purposeful element of the teaching sequence designed to introduce re-sampling. Expressed in terms of the situated abstraction model (Hershkowitz et al., 2001), students demonstrated the first two of the three elements of constructing and recognising mathematical knowledge.

In subsequent activities students re-sampled progressively more independently and with less support from the researcher. In the activity "The effect of sample size on the fairness measure" students modified the sample size, collected measures at three different sample sizes, and contributed their own data to the class measures dot plot, and in the activity "First Fathom coin activity coin toss 50 & 500" students re-sampled independently, collecting multiple measures at two different sample sizes. In both instances, however, the sample size and the number of measures collected were either determined by the researcher or by the number of students physically present in the classroom (one student provided one measure). In the post-study classroom assessment task of the "50 students in a Year 9 maths class," less than half of the students were able to demonstrate the third element of situated abstraction of building-with to transfer and apply their knowledge to a new context. Students appeared to have a clear understanding of a measure within a particular context and could explain the measure by giving an example (Section 4.5.2.5); that is, students had an exemplaric and low level understanding where the word is described by example (Meyer, 2009, p. 910), rather than by definition. Their difficulties arose when the principle of measures was generalised to contextual tasks, such as the short-worded problems given in assessment tasks, which suggested that students did not have a definitional sense of the terms. Students demonstrated elements of constructing and recognizing mathematical knowledge, but students were largely unable to build-with their knowledge. Students were not given the opportunity to choose the number of measures collected, and this may have contributed to their difficulties in distinguishing between the two terms. In the classroom students' written responses to questions of the number of measures were generally correct, but they only needed to refer back to their notes or count the number of measures displayed on a dot plot.

Confounding sample size and the number of measures collected is potentially understandable: both, in a sense, are a sample size. In conventional data collection a sample size refers to the number of people who participated in the survey or more generally the number of data collected. In conventional sampling only one measure is collected – whether it is the proportion in favour, or the proportion of heads, or the number of "sixes." Re-sampling introduces more than one measure collected. The term "number of measures collected" has little meaning except as part of re-sampling. The two expressions were also cumbersome, and it was natural that students reverted to the short-hand expressions of sample size and measures, and this may have exacerbated confounding the two uses. Students' difficulties distinguishing samples and measures occurred principally in short worded tasks where students worked independently. These tasks were highly compressed relative to the classroom activities.

Students' notions of sample prior to the study were largely informal: a sample provided an imperfect representation and highlighted the importance of a representative and a randomly chosen sample. This study extended these largely informal notions to the more formal notion of sample as a measurement with an associated accuracy, explicitly quantifying sample size, and sampling to collect multiple measures. The expressions were a constraint on learning and at an introductory level, such as in this study, it may be more effective to retain a specific title for the measure such as "proportion of heads" rather than use the abstracted and generalised term "measure." This more informal approach, where students use statistical terms loosely, is consistent with an earlier recommendation of Bakker and Gravemeijer (2004). The purpose of generalising the term to measures in the study presented here was to provide students with general re-sampling tools for subsequent study. Students' ideas co-evolved and became progressively more sophisticated in this study, but their understanding was, at best, developing only.

5.5.2 Measures dot plots

The measures dot plot was used in this study to display and to support the analysis of re-sampling data, and consequently students' ability to understand and interpret measures dot plots was considered a skill essential to complete tasks successfully. The purpose of a graph is to draw attention to specific features of the data, and two key re-sampling principles that the measures dot plots were used to promote were that as the sample size increased (a) the centre of the distribution of measures approached the expected value and (b) the spread of the measures decreased. These two principles also provided the conceptual understanding for the large population sample size model, without which the sample model could be used in a procedural sense only. The detailed study extended the classroom study and sought to promote students' reflection on measures dot plots by asking students to compare measures dot plots with the more familiar cumulative proportion of heads graph.

Students' first use of the dot plot format – as distinct from the measures dot plot used in re-sampling – was the pre-test item that examined data of female marathon race times. Few students found the item difficult, and this suggested that students were untroubled by the dot plot format. In this context the dot plot format fulfilled the principle of the less abstracted the better understood (Konold & Higgins, 2003). The data in the dot plot were not abstracted, aggregated, or consolidated: there existed a one-to-one correspondence with the underlying data. If some students found the measures dot plot challenging in the study subsequently, it was most likely not the graph format, but the underlying measures data.

Students had constructed the dot plots over an extended period of several lessons and had contributed their own data, and their analysis was supported by whole-class discussion, the GICS framework that provided a check-list for students to examine the data, and their own notes taken during the class discussion. Students' did not produce work independently, and a familiar task was assessed. The value to learning was principally modelling of the process of statistical analysis.

Measures are derived data, and consequently a measures dot plot is a plot of derived data. In schools students do use graphical representations of partially aggregated data sets such as column or frequency charts. This study used three measures: the fairness measure in the die simulation, the proportion of heads in the coin simulation, and the percentage For in the opinion surveys. In the die simulation students may have formerly considered the frequency with which faces occurred, but this was extended to the fairness measure calculated as the sum of the differences between observed and expected – the fairness measure was an abstracted statistic and an intellectual distance existed between the underlying data and the statistic. Once calculated the direct connection to the underlying data (e.g., the frequency at which each face occurred) was lost and it was not possible to re-construct the original frequency data uniquely from the fairness measure. This abstraction obliged students to shift from additive (consideration of frequency) to the proportional or multiplicative thinking required for the fairness measure. One student described this complexity as "data within data."

To complete the assessment task comparing the three dice using the GICS framework successfully, students were required to shift to the higher level of abstraction of distributional thinking. The Global, Individual, measures of Centre and measures of Spread (GICS) framework was used to encourage students to gather and categorise all the available information before calculating statistics or drawing a conclusion. This task was complex because students considered the three data distributions of the home-made, factory-made and the Fathom die simultaneously. Despite this complexity the students gave sophisticated interpretations of the three measures dot plots, and students' responses were more sophisticated than those reported by Shaughnessy (2006) to similar tasks. Situated abstraction (e.g., Pratt & Noss, 2002) was used to explain how students were able, in this instance, to abstract meaning from the task: the examination of the fairness of the three dice was a highly supported collaborative activity conducted over an extended period of several lessons where students contributed and consequently owned the data (Section 4.5.3.8). The acronym GICS provided a mnemonic and the routine of the framework helped cultivate statistical habits of mind. Introduced and described as a framework, it may be better described as a check-list. Although this could encourage task performance by rote, it did establish a routine and a webbing structure (Noss & Hoyles, 1996) that supported learning to more formal processes. The high level of students' responses suggests a potential pathway to introduce measures dot plots first through dot plots of a familiar data set, and second through a supported task that analyses more complex measures dot plots. This does not, however, demonstrate students' ability to work entirely independently and build-with to apply the concepts to a new context.

Students were less successful subsequently in abstracting meaning from other measures dot plots used in less well supported tasks. For example, few students transferred correctly their knowledge of measures dot plots of the item examining the effect of sample size on the fairness measure (Section 4.4.3) to the coin measures 50 and 500 tosses of a coin homework item – Part 2 (Section 4.4.5) – some students incorrectly displaced the centre of the distribution to the left away from the expected value, which implied that the bias of the coin was affected by sample size. On the post-study assessment students demonstrated an understanding of both, one, or none of the two principles of the effect of sample size on the measures in different contexts. Students' knowledge generally was only partially developed and not robust, and their ability to build-with and transfer their knowledge to a new context was limited.

Students may not have had a strong overall understanding of the data set (Table 4.35), and this lack of understanding may have contributed to some students' difficulties in interpreting measures dot plots and thus been a constraint on learning. In a homework task and on the post-study assessment tasks some students were unable to provide an appropriate name for the measure (Sections 4.5.2.2). An overall understanding of the data would seem essential for sense-making and analysis of any dot plot. A graph may support interpretation of the data, but if the underlying data themselves are not understood robustly, then the graph may not support interpretation (Roth, 1998). Students were also uncertain of the key terminology of sample and measures, and this may have also contributed to their difficulties.

The detailed study, in noting the difficulties students had with measures dot plots in the classroom, extended the classroom study to consider students' examination of the cumulative proportion of heads graph as alternative to the measures dot plot. This graph cued responses of the trend, but students provided a description of such minutiae that the overall perspective was lost, and it took the intervention of the researcher to shift students' attention to specific features of the trend. The trend graph itself did not provoke students' responses to quantify the difference between observed and expected. The cumulative proportion of heads may provide a learning path to the use of measures dot plots, but this was not part of the classroom study. In this study the measures dot plot was introduced and used as a separate, potentially more general tool, than the cumulative proportion of heads graph. The measures dot plot graphs were not displayed explicitly in Fathom, but it could be displayed readily. This approach creates a dilemma. A sophisticated Fathom worksheet that displays explicitly the difference between observed and expected takes time for students to assemble, and a pre-assembled worksheet risks students' not accepting or understanding the simulation's construction or function.

Students in the detailed study recognised the value of the measures dot plot. Students commented that the cumulative proportion of heads graph was easy to understand and explain, and that it contained all the information of an individual coin toss series, but the measures dot plot was described as containing more information, of "data within data," of showing how the data were "clumped." The cumulative proportion of heads graph was familiar to students, but the measures dot plot was novel. Several students expressed the opinion that their actual choice of graph depended on the situation or purpose at hand, which indicated an intelligent choice of the available tools.

The measures dot plot was used as one of the elements of a study introducing the mathematical technique of re-sampling. Students had little difficulty with the dot plot format, but many students found the use of measures dot plots challenging. For these students the study provided only an introduction to measures dot plots that, nevertheless, provided a foundation for more formal study at senior years.

5.5.3 Students' relationship with Fathom

The third and final aspect used to consider Research question 3 and whether Fathom and re-sampling offered an effective learning opportunity for high school students has, as a common theme, students' relationship with Fathom. This was examined through students' procedural use of Fathom, instances where Fathom promoted learning, participants'' attitudes to Fathom, and students' recall of Fathom after a six-week break.

Procedural use of Fathom was defined in Chapter 3 as the basic skills to complete a Fathom task, such as constructing a graph. The objective of using Fathom in the study was to promote learning of the mathematical concepts, not to develop proficiency in the use of the software that would allow students to work entirely independently. This study sought to devote as great a proportion of class time attending to the mathematical concepts as possible and to minimise class time on developing skills to use the software – the software was learnt almost incidentally to the study of the mathematical concepts.. The study adopted a different approach to the preliminary self-instructional courses provided by Biehler (2010), and provided students with the opportunity to acquire skills sufficient for the tasks only. The software was not available for home use and therefore students did not use Fathom independently outside the classroom.

Instrumental genesis provided a mechanism to reflect, from the perspective of tool use, on how the software tool should be introduced and used, and how the software acted upon the user. Instrumental genesis (e.g., Drijvers, Kiernan, & Mariotti, 2010) is the process by which an artefact (a blank Fathom workspace) was combined with schemes to produce an instrument (a Fathom simulation). Within this study one such scheme was the procedural use of Fathom needed to assemble a simulation. Four classroom activities were analysed, of which three were considered successful, and one not as successful. Success in the procedural use of Fathom was demonstrated when students assembled and used simulations reliably, stayed on-task and attended to the activities across a range of learning styles that were inevitably part of any class.

The instances that demonstrated successful implementation of Fathom in this study provide insights into how Fathom could be introduced into the classroom. Students' success was apparently determined by the level of learning support – described in instrumental genesis as orchestration – provided to the students. Examples of successful orchestration included peer instruction with one student, previously instructed by the researcher, acting subsequently as the tutor for a second tutee student, and instruction worksheets that included specific instructions and screen grabs of functioning simulations being assembled. The worksheets were presented as series of screen-capture photographs and specific sequence instructions using arrows that clearly identified steps required to construct the simulation. The screen-capture instructions provided visual cues and a template for constructing the simulation. The detailed worksheets appeared to provide a highly efficient and effective method of introducing the software that allowed students to acquire basic skills and develop acceptance of simulation. The worksheet minimised class time devoted to acquiring procedural skills and to using Fathom while maximising the opportunity to examine the mathematical objectives of the study. At the conclusion of the study students had acquired a basic repertoire of skills that allowed them to assemble die simulations independently. The post-study questionnaire reported that students, particularly the boys, considered the worksheets "the best way to learn how to use Fathom."

In the one instance where the orchestration provided was not successful was the activity examining the effect of sample size on the fairness measure, but it was unsuccessful only in the class where the simplified worksheets that presented a list of single line instruction were used – the same activity with the guided worksheet that used screen-grabs and detailed instructions was successful. The simplified worksheet did not provide sufficient information for students at their stage of development to assemble a functioning simulation.

At the post-study assessment all the male students and 80% of the female students were able to assemble, operate, and interpret a basic simulation independently. In the detailed study six weeks after the classroom study students needed only limited support to use the software, and this suggested that students would be able to use the software in class productively after an extended period when it was not used.

Within the instrumental genesis framework, instrumentation describes the process where the software acts upon the user; instrumentation that acted to support learning was an affordance, and instrumentation that affected learning detrimentally was a constraint. In the detailed study four instances where Fathom acted as an affordance for the user to support their transitions to more sophisticated statistical thinking were identified: (a) frequency to proportional data analysis, (b) language use of "tossing a coin" to "sample size," (c) small to large sample sizes, and (d) interpreting a graph to choosing a graph to "tell a story." Students' shift in preferences from a frequency to a proportional data representation corresponded to a shift from additive to higher-order proportional thinking. This shift was supported by presenting simultaneously displays of both the frequency and proportional representations of the data that allowed students' focus to cycle through the data representations offered.. The shift occurred naturally because students could view the different data representations simultaneously, and at large sample sizes or sample sizes not conveniently divisible by two it was easier to consider proportions of heads rather than the relative frequency of heads. Once the transition to a proportional representation was made and students had opportunities for practice, the proportional representation appeared natural and fluent. Students used the expression "tossing a coin" only with a physical coin: the expression was associated with the act of tossing a coin. With virtual simulation, however, there was no equivalent physical act (other than pressing a key, but that was not unique) so the expression sample size or more simply the numeric value of the sample size was used.. Sample size may strengthen the connection between virtual simulation and a practical application, e.g., to a survey, through use of an expression common to both. Students' notions of sample size were extended by orders of magnitude by virtual simulation, because virtual simulation gave students practical exposure to sample sizes not practicable with physical simulation. After limited use large sample size became common-place, students spoke naturally of sample sizes of several hundred or several thousand. In all three instances of instrumentation Fathom provided the means to change students' thinking directly, and consequently was an affordance to learning.

A fourth example of instrumentation that supported learning was Fathom's graphing feature. Education research has noted that the new software tools must be used in new ways if the power of the software is to be utilised (Ben-Zvi, 2000). One example of using Fathom a new way was the construction of graphs. A traditional pedagogical approach might instruct students to construct a type of graph chosen to highlight specific features of the data. One approach used in this study asked students in the exploratory data analysis of the New York Marathon data set to create a graph using a format of their own choosing with the expressed objective that "the graph must tell a story" (Sections 4.2.5 & 4.5.4.2). Fathom largely eliminated the time involved to construct a graph, and this allowed students to explore a variety of graphs in quick succession as a means of sense-making of the data. This provoked a wide-range of sophistication of students' responses: all students produced a response with meaning to them, and able and engaged students provided rich responses. Fathom also acted as a constraint on learning in two instances. The first was related to managing the display, so that the Fathom workspace was not cluttered, and the second to the default number representation, which led many students to quote an inappropriate number of decimal places.

The Fathom features students used matched the tasks set, which were related to the mathematics accessible to the Year 9 students, and were driven by the class time available and by which software features students could utilise effectively. At the introductory level of this study these tasks were basic data analysis and simulation. Software features, such as, the formula editor, were introduced only as the features were needed for a particular task, no more than two new features were introduced in the one lesson, a consistent approach complemented by a process of progressive and step-wise checking of the simulation, which, for example, included a check that the data representations were internally consistent, was used. Fathom was used to generate individual measures, but the measures were collected and recorded manually because collecting and recording the data entirely with Fathom required learning additional skills with the software. The number of measures collected was typically 30, which was chosen to mimic the number of measures collected by a classroom of students contributing one measure, but such a number did not show clearly that the underlying shape of the distribution approximates a normal distribution.

The extensive range of features within Fathom suggests longevity for the software across high school years. Students will not out-grow the software: the level of functionality would meet the needs of all but a few statistics specialists. Students used the software with minimal instruction, but could continue to learn more powerful features as needed. Fathom had a low entry cost and students were able to use the software productively with minimal instruction. Students in the detailed study had little difficulty using the software after an extended break of six weeks, so its intermittent use in the school year would not disrupt learning. Fathom's longevity, low-entry cost, and apparent ease of use after an extended break would eliminate the need for students to learn new software, and it could reduce the need for schools to buy additional software and to train teaching staff in its use.

A disadvantage of landscape-type software, such as Fathom, over route-type software (Bakker, 2002) is that students may be overwhelmed by complexity of the software or be diverted from task. This did not occur. Students certainly explored the software, but they generally stayed on-task. Students were not confused by Fathom's ability to display several graphs or several data representations simultaneously. The detailed study showed students moving between representations of the same data as they sought to assemble the information to develop meaning.

Fathom offers a modular construction and versatility: the software can be presented to students as black-box / route-type software, white-box / landscape software, or at a point between as "grey-box" software. The term grey-box was introduced to describe the detailed worksheets that provide students with specific instructions to assemble a simulation. It differs from black-box software where a pre-assembled simulation is used, and it differs from white-box software where students would operate independently. Students assembled and checked progressively the simulation and hence developed some appreciation of the construction and operation of the software. Similarly a partially assembled simulation was used in the "Large population sample size" activity where students extended the simulation to include additional features. The modular nature of the software allowed Fathom to be used to generate individual measures, which students recorded and graphed manually – to teach students to use Fathom to collect and graph measures would have required additional tuition and higher abstraction.

The research literature recommended that virtual simulation activities begin with the familiar physical equivalent, such as coins and dice, but this may not be necessary. Students were highly engaged with the activities investigating the fairness of the dice. The second physical simulation, the physical coin, was of dubious value – students could have conceivably used the virtual simulation immediately. The female students were not engaged with an activity that had been studied previously and had no novelty.

Students preferred Fathom over the more readily available MS-Excel™. The student survey identified drag-and-drop, the formula editor, and the method of displaying means as good features. The formatting features in Fathom are less sophisticated than, for example, MS-Excel. Limited formatting features may be an advantage as they reduce the opportunity for students to spend time unproductively in graphic design rather than examining the statistical concepts. Fathom should include a formatting feature to set the decimal place more appropriately. The majority of students believed that they needed additional practice before Fathom could be used effectively. Students and the colleague teacher commented very favourably on the ease of constructing graphs within Fathom.

Students' attitudes to Fathom were explored in the detailed study and the post-study questionnaire. In the detailed study students expected the software to offer advanced features, but in the opinion of the researcher the software's simplicity promoted transparency and use. Students noted that the workspace could be cluttered and specific steps could be momentarily forgotten, but no problem common to many students occurred.

Substantial differences in attitudes to Fathom existed between the male and female students, and the male students considered Fathom more favourably than the female students. This difference in attitude may have been a factor in the female students' low level of submission of work samples. These responses parallel the gender-based difference in attitude to computer-based learning noted in the literature (e.g., Vale & Leder, 2004).

In response to the third and final research question Fathom and re-sampling does offer an effective learning opportunity for high school students. Fathom, used in the manner of the study with guided worksheets, construction of the simulation, systematic checks of the simulation, and modest development of skills in the use of the software appeared to be an affordance to learning. Re-sampling had successful application in the investigation of the fairness of dice, and activities of this type may provide a foundation for the study of the chi-square statistic at more senior levels of education. Re-sampling may allow exploration of sample size and margin of error as an extension of Law of

Large Number activities, and an extension to an application of the large population sample size model to contextual tasks may build on these developing understandings (Section 4.4). Students found use of the general term measures novel and at times a barrier to learning. The use of the two expressions, “sample size used to calculate a measure” and “the number of measures collected”, which were introduced to support the development of generalised re-sampling skills, may have acted as a constraint on learning, but assigning a specific name to the measure such as fairness measure or proportion of heads may be more comprehensible and meaningful to students. The dot plot format was readily understood by students, but the measures dot plot of derived and abstracted data was more challenging. Students were able to interpret measures dot plots in the supported activities in the classroom. Students’ use of Fathom, examined through the aspect of instrumental genesis of instrumentation, supported students’ in their transition from frequency to proportional data representations, from tossing a coin to the more formal expression of sample size, and from small to large sample sizes. Students’ abilities to apply re-sampling independently to contextual tasks were limited and their knowledge could be considered developing only.

5.6 General discussion of the study

This sub-section provides a discussion of aspects of the study not addressed directly in the responses to the three research questions, but which nevertheless are thought important to the themes and the educational content of the study Section 3.3.3.

Whole class discussion and dialogue was identified as an essential element of enculturation into the process of statistical enquiry (e.g., Ben-Zvi & Garfield, 2004). The confidence and ease with which the boys and their colleague teacher spoke demonstrated that class discussion was an integral part of their learning culture, but a culture of discussion was not as evident in the all-girls’ classroom. Class discussion was not – according to one female student – an established part of the classroom practice. Cultivating a culture of discussion within the limited time available for the study was a challenging task. The colleague teacher explained the female students’ reticence largely in terms of students’ lack of familiarity with the researcher, and students’ own lack of confidence and a natural reluctance not to appear foolish.

> It is difficult to promote discussion until the students felt comfortable [….] students are not familiar with you [….] later on [in the study] they were more

> confident, but initially they weren't sure of expectations and whether they should know more. [Colleague teacher of female class]

A key objective of the study was to develop students' ability to communicate their understandings through the development of a statistical vocabulary by means of whole class discussion and written work. Evidence for students' language use was provided by the compare three dice using GICS assessment task, which in itself was a product of the whole-class discussion. Students freely used a variety of formal and informal terms such as "… range, count (taken from Fathom), mean, outliers, clusters, measures of spread and measures of centre (both, apparently, unfamiliar prior to the study), more tightly bunched, either side of the mean, reasonably consistent …" This general use of statistical terms was in addition to the specific terminology discussed earlier in Section 5.5.1. The colleague teacher of the female class noted the difference in pedagogy within the research study emphasising whole-class discussion, written explanation, and the use of the computers. Although recognising the objective of developing intuitive notions of the statistical concepts, the colleague teacher thought that the students also needed more opportunity to practise and develop mathematical sub-skills.

Students may have been uncomfortable with the ambiguity and complexity of statistical analysis. In the post-study student questionnaire Item 22 "maths problems with one clear answer" were preferred by male students in a ratio of approximately 2:1 and female students in a ratio of approximately 3:1. Several students who were procedurally competent in formal mathematics found statistics and probability troubling. There may be potentially three categories of students: students who are procedurally strong and enjoy the precision of mathematics; students who lack confidence in their abilities and prefer the clarity of mathematics; and students who can confidently and competently interpret data, and who are comfortable with ambiguity of statistical analysis.

Some differences in the two classes were observed. It was not clear, however, whether these differences reflected the gender of the two classes, the personal maturity of the students, or simply the different cultures and learning background of the classes and schools, but such differences are consistent with research literature that indicates boys' greater interest in computer-based learning and girls' relatively greater interest in contextual tasks (Vale & Leder, 2004) A general observation from this study was that the male students were mathematically procedurally stronger and more enthusiastic participants in the physical activities than the female students. The female students gave more sophisticated responses to the tasks than the males, and this is consistent with literature that females have stronger language skills. If the differences in the behaviour between the two classes was indeed related to gender then it may be more effective to offer computer-based learning opportunities tailored to male and female students' learning styles.

An inconsistency within the girls' advanced mathematics class was the relatively low level of numeracy skills amongst a proportion of the female students (Section 4.2.1). For example, even if it was inappropriate to do so, both girls and boys often continued to use extended decimal values (Section 4.2.5). This showed a lack of understanding of the data set, a lack of number sense about the data, or prior experiences where this type of response was expected. It was also not clear whether students had a strong sense of percentage, for example, one student described 46% as a majority.

Students were exposed to learning experiences that extended far beyond the development of mathematical concepts. For example, a colleague teacher and a school principal noted that much of the benefit of this study lay with exposing students both to the university and to mathematics education research. Students also had the potential of contributing to the development of international software through a direct live video contact with the principal Fathom software developer. Other experiences were more general. Developing statistical habits of mind is a process of cultural and attitudinal change, and that evolutionary process defies convenient precise measurement. Within the activities students were provided with the experiences of statistical enquiry, simulation, modelling, and practical application of mathematics.

The school generously donated colleague teachers' and students' time and school resources. Conducting extensive research in schools is problematic: practising professionals who have an established profile in the school do not have time to conduct research, and researchers outside the school do not have an established presence and profile in the school. A long-term collaborative approach is essential, but maintaining and cultivating the relationship requires involvement at the university and departmental system level. The importance of developing a long-term relationship with students, teachers, and schools to support education research is discussed elsewhere (Bill, 2010).

Item 30 on the post-study questionnaire asked students whether they considered the unit of work worthwhile. The item was worded purposefully. It was designed to encourage students to reflect on the experience, rather than to respond at the emotional level of whether they enjoyed the work. The boys were overwhelmingly positive, with only one student giving a neutral response and no boy responding that the work was not worthwhile. The girls were not as strongly positive, but almost two-thirds of females (62.9%) responded that the unit was worthwhile and 20% were neutral. A proportion of females, 17.1%, however, did not think the unit was worthwhile, but this perception was not explored in the study.

5.7 Limitations of the study

This sub-section considers the choice of student cohort as providing data representative of the broader student community, the topic as appropriate for Year 9 students, the effect on students' learning of the short time frame of the study, the limitations of the instrumental genesis and situated abstraction frameworks as applied in the study, and the role of the industry partners and the colleague teachers.

The extended mathematics classes, with many capable and motivated students, were not truly representative of the wider high school population. The students were eager to complete the pre-test and demonstrated generally a very good level of on-task behaviour, but in common with any class the level of motivation fluctuated. Students' interest appeared to decline and a more focused study, with Fathom used over a longer period and incorporated into other aspects of mathematics and other subjects, for example science, may be more effective.

The basic skills pre-test showed the students had the computational and conceptual skills required to complete the activities and tasks within the study. The post-study interview revealed that one colleague teacher considered the topics entirely appropriate for the class, and the other colleague teacher was supportive of the topic as suitable for the students. The teachers were divided on whether the large population sample size model $e = \pm 1/\sqrt{n}$, was appropriate for the Year 9 class. The colleague teacher of the boys' class thought the topic was appropriate and a natural extension of random and representative sampling, but that the use of measures was better suited to the subsequent Year 10 course. This teacher supported the introduction and use of the sample size model, but also recognised the range of abilities present in the class. The colleague teacher of the girls' class thought it was important to include practical contextual tasks such as the sample size problems considered in the study. In response to whether use of the most advanced concepts of the sample size model was too complex for the students, the colleague teachers said:

> No, not at all. This was a highly spread class. The more able students would have grasped the concept, but in this class only about a half got a handle on it. Other students perhaps developed a feel for the concept and took away a general picture of how many to sample, even if they could not give a specific answer of how many to sample. [Colleague teacher of male class]
>
> There was a whole range of students. For some of them I felt they got the big picture, some who got most, others were doing the mechanics of what you wanted them to do without really understanding quite why they were doing it and how it fitted in. I suppose about a quarter [of the class] did [the mechanics] and didn't quite get the big picture. [Colleague teacher of female class]

The research study was conducted over a short three-week unit of work, and this may be insufficient time to change student thinking permanently. The study demonstrated, in response to Research question 1, students' acceptance of the Fathom simulation as legitimate, but it did not demonstrate significant permanent change in students' understanding of sample size (Research question 2). A colleague teacher noted that most teachers would expect that learning occurs most productively on any concept over a sustained period of time. Students needed time to be confident users of Fathom, and time to be confident users of all the skills developed. There was insufficient class time to explore adequately the complex concept of sample size. The cognitive load on the students was also high, and this load may have affected their learning. Within the study students needed simultaneously to acquire basic procedural skills in the software, develop understanding of new mathematical concepts, and adapt to an unfamiliar teacher and pedagogy. The tasks became progressively more complex and less supported by the researcher and the colleague teachers during the research study. The change in both task complexity and support obscured the longitudinal development of the students and complicated the study. The study may have been more effective if conducted in two studies separated by several weeks or months, with the first addressing students' acceptance of the simulation and the second study investigating sample size. The two participating schools was very generous in agreeing to participate in the study, but two separate studies or an additional time allocation in an already crowded curriculum would not have been a reasonable request to make.

The study examined samples and sampling, data distributions – some of the big ideas of statistics. The study had the resources, time, and opportunity to conduct research using the one teaching sequence only, but limiting research on the important big statistical ideas to the one teaching sequence does the topics a disservice. The topics warrant further research using a variety of teaching approaches.

Instrumental genesis (e.g., Drijvers, Kiernan, & Mariotti, 2010), supported by aspects of situated abstraction (e.g., Pratt & Noss, 2002) gave a means to observe student learning, but neither is supported by a formal framework or procedure. The analysis of the data was highly subjective and open to alternative interpretation. There was scope within the thesis to examine selected aspects only, chosen by the researcher as significant, but such a choice is also subjective. Despite these qualifications instrumental genesis and situated abstraction were used to inform the research and complement the broader perspective offered by the SOLO assessment of the classroom study.

The key stake-holders of the two industry partners Key Curriculum Press and the Australian Bureau of Statistics, the University of Tasmania, the Tasmanian Department of Education and the two participating schools placed few constraints on the research beyond requiring that all ethical responsibilities were met. The researcher felt a strong professional obligation to provide research of value to the industry partners. It was, however, a challenge to identify a research topic that was practicable and worthy of research. There was an entirely legitimate desire of the software developers to extend the boundaries of their knowledge of software use in schools. In this study, however, the software was not already established in schools and the students used the software in the most elementary way, so the potential to provide information for the software developers to extend their knowledge of the use of the software was limited. The colleague teachers were supportive, but given the commitments as senior teachers they declined the offer to develop basic skills in the software or participate substantially in the design of the study, and consequently, the colleague teachers' ability to support students and to provide substantial feedback on the teaching unit was limited.

5.8 Implications of the study

This sub-section considers the implications of the study for teachers, statistics education researchers, software developers, and teaching resource developers.

5.8.1 Implications for teachers

- Incorporate measurement accuracy as an element of physical measurement in the mathematics classroom. Neither of the two participating schools appeared to have considered the accuracy of ordinary physical measurement of length, mass, or time extensively and students had little concept of measurement error. An understanding of the error or tolerance and the practical significance of measurement error is part of the cultivation of number sense and sense-making, and a prerequisite to consider the accuracy of random sampling.

- Purposefully support students' acceptance of the ambiguity of probability and statistics. An attraction of traditional mathematics to students – particularly the male students in the study – was that formal mathematics offers definitive answers: students knew when they had obtained a correct answer and felt rewarded when they did so.
- Use the GICS framework to support students' verbal analysis of statistical data. The GICS framework – essentially a check-list – used in the study was designed to encourage students to note the global, individual, centre, and spread features of a data representation, and then assemble them into a coherent verbal analysis of the data. This framework may have supported students' sophisticated responses in the analysis of the fairness of the three dice.
- Introduce the determination of sample size into the high school curriculum. The school curriculum emphasises random and representative sampling, but it does not include the natural complement of a consideration of sample size. Virtual simulation may provide a tool to develop intuitive understanding of sample size, and a means to explore small populations likely to be familiar to students, such as school size populations and the infinite populations of coin and die simulations.
- Introduce consideration of sample size as a formal mathematical extension of Law of Large Numbers activities. The large population sample size model $e = \pm 1/\sqrt{n}$ could be meaningfully introduced to formally quantify the difference between observed and expected values. The sample size model has no practical value when used with physical simulation because of the small sample sizes practicable with coin tosses and die rolls, but virtual simulation allows the model to be used with large sample sizes that allow simple calculations.
- Consider developing students' intuitive sense of sample size in small populations before examining large or infinite populations. The model in the

study calculated the sample size for large or infinite populations, and it led to sample sizes larger than the populations familiar to students. One student wrote that the 10% rule can be applied in all practical sampling situations, which carries the implication: why learn more than one model? This study did not provide students with a sample size strategy for small populations, and this may have been a factor in the persistence of the 10% model.

- Use Fathom in high schools. The software has a relatively low entry-cost, it is versatile, and it has features that will ensure its longevity across the school years. The software has widespread application across a range of students' ages and abilities. Fathom is a powerful tool. Fathom's low cognitive entry cost was demonstrated by the impressive responses of the students in the exploratory data analysis task in which all students produced good results with minimal teacher instruction, Fathom's versatility was shown by the use of its exploratory statistical analysis and simulation features by students with a range of abilities, and Fathom's longevity is shown by the formula list that offers features well in advance of that expected for high school students. The software also provided alternative learning opportunities and pathways, and these opportunities, particularly for less able students, could contribute to the development of students' confidence in mathematics. At present, given the highly constrained syllabus at senior Years 11-12, the introduction of Fathom in Tasmania would be difficult, but Fathom could be introduced in Year 10 to develop skills in anticipation that the software would be used more extensively in senior years.

- Present Fathom to students as a range of white-box, black-box, or, as in this study, grey-box software. When used as a white-box tool Fathom lent itself readily to exploratory data activities and the transparency supported students' acceptance of simulations. Students used the software productively with limited instruction. Students also had little difficulty recalling the use of basic features of Fathom several weeks after the unit of work was completed, and this suggests that the software could be used intermittently during the school year. Using Fathom as black-box route-type software when students have accepted simulations allows students to focus on the underlying mathematical content rather than manipulating the software tool. As students' interest in assembling the simulations waned during the study, they seemed more accepting of pre-

prepared Fathom simulations. Grey-box use provides a point between where students attend to some, but not all, of the procedural tasks associated with software use.

- Use detailed instruction sheets to assemble Fathom simulations. The worksheets that included screen-capture graphics and provided explicit instructions were effective: the students appeared to enjoy constructing a functioning simulation successfully and remained on task, which liberated the researcher to support students' development of the mathematical concepts. Assembling and checking the operation of the simulation served the additional purpose of encouraging students to think more deeply about the simulation. Such support could be phased out progressively. When a simplified worksheet was used prematurely, few students constructed the simulation independently and the consequence was off-task behaviour, and an unproductive and unsuccessful class; in this instance Fathom was a constraint on learning. As the study progressed the novelty of using Fathom faded and students' attention shifted to the mathematical tasks at hand – in instrumental genesis terms the tool was being incorporated into students' unconscious thought. Purposefully developing proficiency in the use of the software seems misdirected effort when this proficiency will be acquired incidentally as students use the software, and in this study the software was used productively without the formal preliminary training provided by other researchers (e.g., Biehler & Prommel, 2010). Students are likely to use more widely available software such as MS-Excel outside of school, so developing a working proficiency or fluency in a variety of software packages has parallels with, for example, many Europeans having a working knowledge of several languages.

- Use Fathom to reverse the conventional sequence of graphical analysis. When a graph is constructed manually the type of graph is chosen before the graph is constructed. Fathom eliminates the labour of graph construction and allows the user to create a graph first and then consider whether the graph displays the information effectively and persuasively.

- Introduce students to the sample size model using coin simulations, and defer application of the model to contextual survey tasks to subsequent school years. The application of the sample size model to large population contextual tasks in this study raised separate challenges because it required the transfer of techniques to an entirely different context. For example, the concept of having a fixed opinion and being chosen at random in a survey was interpreted by some students incorrectly as expressing their opinion at random or chaotically. Students' erroneous beliefs of sample size were also extremely persistent, and the limited time spent on applying the model to contextual tasks confounded many students.

- Use simpler alternatives to the expressions "the sample size used to calculate a measure" and "the number of measures collected" when introducing re-sampling. The measure, a synonym for a statistic, might be introduced specifically by name, for example, "proportion of heads," rather than generalised to measures.

- Give great emphasis in formal mathematics to the study of statistics and probability in schools. Much of the existing statistical education research in schools focuses on building students' intuitions, but such an emphasis, while essential, is not sufficient because it does not adequately recognise that the mathematics curriculum lies on a path to an increasingly formal mathematical approach at senior school. A failure to provide opportunities for more formal mathematising has at least two consequences: it does not provide a developmental pathway for more formal mathematical study, and it does not maximise the opportunities for exposure to novel mathematics that may extend students' knowledge. The two formal mathematising processes introduced during the study were the development of the fairness measure and the large population sample size model, and the study also provided opportunities to practice sub-skills including calculations involving fractions and decimals.

5.8.2 Implications for statistics education researchers

This study raised far more questions than it answered, and these questions provide the basis for further research. The following are proposed as topics for further research.

- Examine pathways to develop students' intuitive understanding of quantifying sample size. The decision to explore large populations was based principally on the belief that the mathematics involved – the large population sample size model – was potentially more accessible to students than small population sample models, but small populations of 1000 or fewer that are encountered in school are likely to be more intuitive and natural for students.

- Examine students' intuitive sense of the relationship between sample size and population size. In the study students demonstrated inconsistent beliefs of sample size: students were seemingly content to use a sample size of fewer than 100 for an infinite population of a coin toss and a sample size of 15,000 for a large but finite population.
- Compare students' intuitive sense of the frequency distribution of a coin toss with their intuitive sense of the distribution of the proportion of heads from a coin toss. The pre-test examined students' sense of the proportion of heads in 50 tosses of a coin, but students did not find the task intuitive and a substantial proportion of the male students were not confident to provide a response. Students' intuitive sense of the frequency of heads of a coin toss has been studied extensively (e.g., Shaughnessy, 2007), but not with the data presented as proportion of heads. Using these items as a paired task with data presented as both frequency and as the proportions of heads could conceivably yield entirely different learning outcomes. The items lend themselves readily to exploration using simulation, and the item provides a basis for the informal exploration of the binomial distribution, the central limit theorem, and measures of data spread.
- Examine students' sense of sample size when sampling directly from very large populations. In the study large populations were modelled indirectly using infinite populations because Fathom was not able to support very large populations directly (where each individual is actually present in the data set). A large population represented as a data set of, for example 50,000 individual data points sampled randomly, may be simulated but the simulation operates too slowly to be practicable. The use of large populations, such as 50,000 or larger, may be possible as increased computer power and speed become available.
- Consider introducing students to sample size models by sampling from small populations first, and then extending sample size to large populations. Students had a strong intuitive sense that a sample must represent a proportion of the population; this intuition is correct, but only for small samples. The small population sample model (Appendix H.3) recognises the intuition of the sample

as a proportion of the population by including a factor in the sample model of the sample as a proportion of the population. It is only for a large population that this factor approaches unity and is eliminated, and the small population model collapses to the simpler large population model. Rather than directly confront students' intuition of a sample as a significant proportion of the population, a more productive approach may be to accept the intuition and demonstrate that the influence of sample as a proportion diminishes and becomes inconsequential as the population size increases. Investigating sampling from small populations acknowledges students' intuitions of the sample as a proportion of the population. This approach does, however, have the disadvantage that the sample model is complex and students may have difficulty even using the model in a simple procedural sense of substituting numeric values into the model.

- Explore whether students consider a computer simulation as a legitimate model of a contextual task. Students accepted the computer dice as fair, but that does not imply that students considered a computer simulation of contextual sampling tasks, such as an opinion survey, as a legitimate model of the context.

5.8.3 Implications for Fathom software developers

The study identified several modification and additional features that the Fathom software developers may wish to consider including when up-grading the software.

- Incorporate a line or arrow drawing tool in Fathom. A line or arrow drawing tool, which may be used to draw attention to specific features in a Fathom document, has two ready applications. The first is as a teaching aid when the teacher has access to a data projector and can project the image of the Fathom file to the class, and the second is to allow students to demonstrate their understanding more easily by highlighting specific features of the Fathom workspace or data.
- Present all graph formats, but grey-out options that are not available for a particular data type. Fathom defaults attributes to be either numeric or categorical (e.g., nationality), and depending upon this classification, determines what types of graph are available and what statistics may be calculated. Some attributes may be categorised as either numeric or categorical, for example "year" or "the face of a die," and the default classification left students

wondering why the preferred graph was not available. The default setting can be over-ridden readily, but this is an additional complication for novice users. A possible alternative may be to have all graph formats presented, but some greyed-out and unavailable.

- Provide a conventional time format, or a feature to convert digital time to the more conventional analogue time. Fathom's default digital hours time format may be potentially confusing for students.
- Provide data sets of interest and relevance to the user that are totally compatible with Fathom (e.g., Watson, Beswick, Brown, Callingham, Muir, & Wright, 2011). Many of the data sets provided with Fathom were developed for the North American education community, and these data may have limited appeal for Australian students. The data sets also use imperial units, and Australian students are accustomed to the metric system.
- Provide, or allow generation of, very large population collections (e.g., 200,000) from which a large sample derived data set (e.g., 3,600) may be taken. In Fathom large populations are presently modelled by infinite populations, such as a coin and die systems, but this introduces the abstraction that the large population cannot be viewed directly. The availability of large population collections allows the large sample size model to be directly linked to the population. Small populations can be modelled in Fathom explicitly, but the more complex small population sample size model must be used. The computing power of personal computers available presently may limit the size of data sets.

5.8.4 Implications for Fathom teaching resource developers

Fathom was designed for senior high school, college and tertiary level, and at Year 9 and in this study only a fraction of the power and versatility of Fathom was utilised. The introduction of Fathom into high school provides students with an opportunity to develop basic familiarity with the software that provides a foundation for the use of Fathom at more senior levels. The sophistication of the software becomes a constraint on learning only when students are presented with software that is too complex, and offers too many options or too many steps, with an unintuitive layout where students become lost. Teaching resource developers are encouraged to include the most elementary use of the software.

- Develop black-line teaching resources to construct simulations that provide clear unambiguous instructions that include screen grabs and step-by-step instructions. The detailed worksheets developed for this study allowed students to construct the simulations with little difficulty, provided students with the satisfaction of constructing the simulation, guided students to check the functioning and the internal consistency of the data representation, supported ownership and acceptance of the simulation, and obliged students to slow and examine the operation of the simulation.
- Develop teaching resources that enable teachers to explore key fundamental principles using a variety of approaches. Providing a variety of approaches creates opportunities to reinforce the principles, practice sub-skills, and provide alternative learning pathways.
- Publish teaching resources only when the resource has proven and documented effectiveness in the classroom. The software developers have published resources to support the use of the software (Erickson, 2008; Key Curriculum Press, 2007), but research evidence supporting these resources is not yet published. Unproven resources should be published purposefully in limited circulation as Beta-test versions.
- Provide teaching resources developed specifically for secondary school that support the teaching of simple probability and statistical principles. The teaching resources presently available examine sophisticated statistical principles, and only some of these resources are suitable for high school. The inclusion of teaching material suitable for more senior students obliges teachers to be more selective regarding what activities are used, and the resource may not be seen as value for money spent because only part of the resource can be used. It is likely that a wider range of resources suitable specifically for high school will become more readily available as Fathom becomes more widely used.
- Promote the use of Fathom across the spectrum from white-box to black-box and points in between as grey-box mode. The study provided some evidence that the software may be used effectively when students assembled the simulations (white-box), and this supported students' acceptance of the

software. The novelty waned and students' interest in assembling the simulations declined, and from that stage students readily made minor modifications to the simulation (grey-box) rather than entirely assembling the simulations. A logical extension of this approach to the use of Fathom is to use entirely in black-box mode as applets with the simulation provided entirely as fully operating software.

5.9 Summary

This study investigated the use of Fathom statistics education software in two Year 9 classes in two high schools in Tasmania, Australia. The study used statistics education best-practice principles that included the cultivation of statistical practice and enquiry. Fathom statistics software, consistent with statistics education research, played a supporting role. Simulation and re-sampling are used widely in education: in junior schools to develop intuitive understanding, and at senior school and tertiary level to support the development of formal mathematics. This study adopted an approach somewhat midway between these two approaches and that included formal mathematics considered appropriate for high school students.

The study examined three research questions, and the first Research question considered students' acceptance of the legitimacy of a Fathom die simulation as a proxy for Fathom simulations generally. This was examined in terms accessible to Year 9 students of whether the Fathom die was fair. The investigation of the fairness of the die simulation also served the dual purpose of providing an opportunity for students to participate in the process of statistical enquiry. The study found that students accepted the Fathom simulation as a legitimate mathematics tool, such that acceptance was not a barrier to learning. The fairness measure, the statistic developed to examine the fairness of a die, was found to be mathematically accessible for students.

The second Research question examined whether the large population sample model $e = \pm 1/\sqrt{n}$ was accessible to high school students. Here the study was inconclusive. Students on the post-study assessment, particularly the females, provided sophisticated responses to a contextual sample size task of a public opinion survey, but a follow-up test item conducted two months after the conclusion of the classroom study indicated that students' long-term development of understanding of sample size was modest, and their beliefs had reverted largely to those held prior to the study. Students generally found the mathematics associated with the model challenging.

The third Research question examined whether Fathom simulation and re-sampling provided an effective learning opportunity for high school students. The study used Fathom re-sampling to investigate the fairness of a die using a formal statistic of the fairness measure and a frequentist approach to demonstrate the large population sample size model $e = \pm 1/\sqrt{n}$. The study examined three aspects peculiar to re-sampling of the terminology used, graphical representations of re-sampled data described in the study as measures dot plots, and students' relationship with Fathom. Students found the two key sampling expressions, the "sample size used to calculate a measure" and "the number of measures collected" confounding. Students had no difficulty in using dot plots of familiar data sets such as New York marathon race times, and little difficulty in using measures dot plots when a direct association with the data existed, but found higher levels of abstraction with less supported tasks challenging. Students were provided only with sufficient procedural skills to use the software to complete the task at hand using instruction worksheets that included screen grabs of a Fathom simulations and detailed instruction. This approach was effective: students assembled simulations successfully, accepted the Fathom simulation as fair, and by the conclusion of the study had acquired sufficient skills to assemble a basic simulation independently. Fathom appeared to support students' transitions from frequency to proportional data analysis, from the language of tossing a coin to sample size, from small to large sample sizes, and from interpreting a graph to choosing a graph independently that "told a story."

The explicit determination of sample size remains an important, fertile, and largely unexplored, topic for education research in high school, and this study has made only the first steps to examine the associated concepts in high school. The large population sample size model has considerable potential as a formal mathematical extension to Law of Large numbers activities presently used in schools by providing an accessible estimate of the relationship between sample size and a 95% confidence interval. Fathom, and the readily availability of other virtual sampling tools may provide a means of investigating the concepts in schools and developing students' intuitions of sample size.

References

Abrahamson, D. (2006). The shape of things to come: The computational pictograph as a bridge from combinatorial space to outcome distribution. *International Journal of Computers for Mathematical Learning, 11*(1), 137–146. Retrieved August 12, 2009, from http://www.springerlink.com/content/qq4021456um5066v/fulltext.pdf

Abrahamson, D. (2009). Orchestrating semiotic leaps from tacit to cultural quantitative reasoning – The case of anticipating experimental outcomes of a quasi-binomial random generator. *Cognition and Instruction, 27*(3), 175–224. Retrieved August 11, 2009, from http://dx/doi.org/10.1080/07370000903014261

Abrahamson, D., & Cendak, R. M. (2006). The odds of understanding the law of large numbers: a design for grounding intuitive probability in combinatorial analysis. In J. Novotná, H. H. Moraová, M. Krátká, & N. Stehlíková (Eds.), *Proceedings of the 30th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 1–8). Charles University, Prague: PME. Retrieved January 11, 2011, from ttp://gse.berkeley.edu/faculty/dabrahamson/publications/abrahamson-cendak_pme30.pdf

Ainley, J. (2000). Transparency in graphs and graphing tasks: An iterative design process. *Journal of Mathematical Behaviour*, *19*, 365–384. Retrieved May 14, 2008, from http://www.le.ac.uk/se/currentstaff/janetainley/documents/Transparency.pdf

Albert, J. (2006). Interpreting probabilities and teaching the subjective viewpoint. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth Yearbook* (pp. 417–433). Reston, VA: National Council of Teachers of Mathematics.

Albert, J. H. (2003). College students' conceptions of probability. *The American Statistician*, *57*(1), 37–45. Retrieved Oct. 13, 2009, from http://www.jstor.org/stable/pdfplus/3087276.pdf

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives.* With P.W. Airasian, K. A. Cruikshank, R. E. Mayer, P.R. Pintrich, J. Raths, & M. C. Wittrock. New York: Longman.

Anthony, G., & Walshaw, M. (2007). Characteristics of effective pedagogy for mathematics education. In H. Forgasz, A. Barkatas, A. Bishop, B. Clarke, S. Keast, W. T. Seah, et al. (Eds.), *Research in Mathematics Education in Australasia 2004-2007* (pp. 195–222). Rotterdam, the Netherlands: SENSE Publishers.

Archbald, D., & Newmann, F. (1988). *Beyond standardised testing: Authentic academic assessment in the secondary school*. Reston, VA: NASSP Publications.

Artigue, M. (2000). Instrumentation issues and integration of computer technologies into secondary mathematics teaching. *Proceedings of the Annual General Meeting of the GDM. Potsdam, 2000.* Retrieved December 12, 2010, from http://webdoc.sub.gwdg.de/ebook/e/gdm/2000/artigue_2000.pdf

Artigue, M. (2002). Learning mathematics in a CAS environment: The genesis of reflection about instrumentation and the dialectics between technical and conceptual work. *International Journal of Computers for Mathematical Learning, 7*, 245–274. Retrieved January 17, 2011, from http://www.springerlink.com.ezproxy.utas.edu.au/ content/ l37tw62107077507/fulltext.pdf

Australian Bureau of Statistics. (2009). *CensusAtSchool.* Retrieved January 8, 2010, from http://www.abs.gov.au/websitedbs/cashome.nsf/Home/Home

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (n.d.). *My School* website (for all-boys' school participating in study). Retrieved August 12, 2011, from http://www.myschool.edu.au/MainPages/SchoolProfileRep.aspx?SDRSchoolId=610000000228&DEEWRId=8070&CalendarYear=2010&RefId=gzH6Bdi32Fe3LC6KrEJeKBimI2PEsEPm

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (n.d.). *My School* website (for all-girls' school participating in study). Retrieved August 12, 2011, from http://www.myschool.edu.au/MainPages/SchoolProfileRep.aspx?SDRSchoolId=610000000237&DEEWRId=6786&CalendarYear=2010&RefId=gzH6Bdi32Ffc45qoaKQaQz%2foT8weCBp1

Australian Curriculum, Assessment and Reporting Authority [ACARA]. (2009). *Shape of the Australian curriculum: Mathematics.* Retrieved August 1, 2011 from http://www.acara.edu.au/verve/_resources/Australian_Curriculum_-_Maths.pdf

Australian Education Council. (1994). *Mathematics – A curriculum profile for Australian schools*. Carlton, Vic.: Curriculum Corporation.

Bakker, A. (2002). Route-type and landscape-type software for learning statistical data analysis. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on the Teaching of Statistics (ICOTS6), Cape Town, South Africa*. Voorburg, The Netherlands: International Statistics Institute. [CD-ROM]

Bakker, A. (2004). *Design research in statistics education: on symbolizing and computer tools*. Unpublished PhD thesis, Universiteit Utrecht, The Netherlands. Retrieved August 10, 2009, from http://igitur-archive.library.uu.nl/dissertations/2004-0513-153943/inhoud.htm

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Batanero, C., & Diaz, C. (2007). The meaning and understanding of mathematics. In K. Francois & J. P. van Bendegem (Eds.), *Philosophical dimensions in mathematics education: Vol. 42*(5). *Mathematics Education Library* (pp. 105–127). New York, NY: Springer.

Batanero, C., Henry, M., & Parzysz, B. (2005). The nature of chance and probability. In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 15–37). New York, NY: Springer.

Batanero, C., & Sanchez, E. (2005). What is the nature of high school students' conceptions and misconceptions about probability? In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 241–266). New York, NY: Springer.

Batanero, C., & Serrano, L. (1999). The meaning of randomness for secondary students. *Journal for Research in Mathematics Education, 30*(5), 558–567. Retrieved January 20, 2011, from http://www.jstor.org.ezproxy.utas.edu.au/stable/pdfplus/749774.pdf

Ben-Zvi, D. (2000). Towards understanding technological tools in statistical learning. *Mathematical Thinking and Learning, 2*(1&2), 127–155. Retrieved November 18, 2009, from http://dx.doi.org/10.1207/s15327833mtl0202_6

Ben-Zvi, D. (2004a). Reasoning about variability in comparing distributions. *Statistics Education Research Journal, 3*(2), 42–63. Retrieved January 10, 2011, from http://www.stat.auckland.ac.nz/~iase/serj/serj3(2)_benzvi.pdf

Ben-Zvi, D. (2004b). Reasoning about data analysis. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 121–168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45*(1), 35–65.

Ben-Zvi, D., & Friedlander, A. (1997). Statistical thinking in a technological environment. In J. Garfield & G. Burrill (Eds.), *Research in the Role of Technology in Teaching and Learning Statistics* (pp. 45–55). Voorburg, The Netherlands: International Statistical Institute. Retrieved July 27, 2011, from http://sites.google.com/site/danibenzvi/Ben-ZviandFriedlander1997Statisitica.pdf

Ben-Zvi, D., & Garfield, J. (2004). Research on reasoning about variability: A forward. In D. Ben-Zvi and J. Garfield (Guest Eds.), Reasoning about variability. *Statistics Education Research Journal, 3*(2), 4–6. Retrieved June 12, 2006, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_forward.pdf

Ben-Zvi, D., Garfield, J., & Zieffler, A. (2006). In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth Yearbook.* (pp. 467–481). Reston, VA: National Council of Teachers of Mathematics.

Berg, C. A., & Phillips, D. G. (1994). An investigation of the relationship between logical thinking structures and the ability to construct and interpret line graphs. *Journal of Research in Science Teaching, 31*(4), 323–344. Retrieved June 20, 2010, from http://onlinelibrary.wiley.com/doi/10.1002/tea.3660310404/pdf

Biehler, R. (1994). "Probabilistic thinking, statistical reasoning, and the search for causes – Do we need a probabilistic revolution after we have taught data analysis?" In J. Garfield (Ed.) ICOTS4 (pp. 20–37). Minneapolis: University of Minnesota.

Biehler, R. (1997). Software for learning and doing statistics. *International Statistical Review, 65*(2), 167–189. Retrieved August 1, 2011, from http://www.jstor.org/stable/pdfplus/1403342.pdf?acceptTC=true

Biehler, R. (2003). Interrelated learning and working environments for supporting the use of computer tools in introductory classes. *Proceedings of the IASE satellite conference on Statistics Education and the Internet, Max-Planck-Institute for Human Development, Berlin.* Voorburg: Netherlands: International Statistical Institute. Retrieved February 14, 2007, from www.stat.auckland.ac.nz/~iase/publications/6/biehler.pdf

Biehler, R. (2006). Working styles and obstacles: Computer supported collaborative learning in statistics. In B. Chance & A. Rossman (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS7), Salvador, Bahia, Brazil.*Voorburg, The Netherlands: International Statistical Institute. Retrieved January 11, 2011, from http://www.stat.auckland.ac.nz/~iase/publications/17/2d2_bieh.pdf

Biehler, R., & Prommel, A. (2010). Developing students' computer-supported simulation and modelling competencies by means of carefully designed working environments. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistics Institute. Retrieved October 15, 2010, from http://www.stat.auckland.ac.nz/~iase/publications/icots8/icots8_8d3_biehler.pdf

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning. The SOLO taxonomy (Structure of the Observed Learning Outcome).* New York: Academic Press, Inc.

Bill, A. F. (2006). *Teaching distribution in a Fathomable™ way.* Unpublished honour's thesis, University of Tasmania, Hobart, Australia.

Bill, A. F. (2010). Researchers cultivating a long-term relationship with schools. In C. Reading (Ed.), *Data and Context in statistics education: towards an evidence based society. Proceedings of the Eight International Conference on Teaching Statistics. (ICOTS8, July 2010). Ljubljana, Slovenia.*Voorburg, The Netherlands: International Statistics Institute.

Bill, A. F., Henderson, S., & Penman, J. (2010). Two test items to explore high school students' beliefs of sample size when sampling from large populations. In L. Sparrow, B. Kissane & C. Hurst (Eds.), *Shaping the future of mathematics education.* (Proceedings of the thirty-third annual conference of the Mathematics Education Research Group of Australasia, Vol. 1, pp. 77–84). Fremantle, WA: MERGA. Retrieved January 12, 2011, from http://www.merga.net.au/publications/counter.php?pub=pub_conf&id=833

Blythe, T. (1998). *The teaching for understanding guide / Tina Blythe and the teachers and researchers of the Teaching for Understanding Project.* San Francisco: Jossey-Bass Publishers.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42. Retrieved July 28, 2011, from http://www.jstor.org/stable/pdfplus/1176008.pdf?acceptTC=true

Brown, J., Stillman, G., & Herbert, S. (2004). Can the notions of affordance be of use in the design of a technology enriched mathematics curriculum? In I. Putt, R. Faragher & M. McLean (Eds.), *Mathematics education for the third millennium: Towards 2010* (Proceedings of the twenty-seventh annual conference of the Mathematics Education Research Group of Australasia [MERGA], Townsville, Vol. 1, pp. 119–126). Sydney, NSW: MERGA. Retrieved December 8, 2010, from http://www.merga.net.au/publications/counter.php?pub=pub_conf&id=205

Buchberger, B. (1989). Should students learn integration rules? *SIGSAM Bulletin, 24*(1), 10–17. Retrieved August 15, 2011, from http://www.risc.jku.at/publications/download/risc_350/1990-00-00-A.pdf

Burke Johnson, R. (2009). Towards a more inclusive "Scientific research in education" *Educational Researcher, 38*(6), 449–457. Retrieved August 19, 2010, from http://proquest.umi.com/pqdweb?index=15&did=1858715301&srchmode=1&sid=1&fmt=6&vinst=prod&vtype=pqd&rqt=309&vname=pqd&ts=1282189405&clientid=20931

Campbell, K. J., Watson, J. M., & Collis, K. F. (1992). Volume measurement and intellectual development. *Journal of Structural Learning and Intelligent Systems, 11*(3), 279–298.

Carmichael, C., & Hay, I. (2009). Gender differences in middle school students' in a statistical literacy context. In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divided: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australisia* (Vol. 1). Palmerston North, NZ: MERGA. Retrieved August 2, 2012, from http://www.merga.net.au/documents/Carmichael2_RP09.pdf

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and perceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75–100. [Electronic version] Retrieved August 10, 2009, from http://ft.csa.com/ids70/resolver.php?sessid=bg9tqe25qulis4lpf1l6hi5r10&server=csaweb107v.csa.com&check=d081173c7a059cd24448ebff6414012e&db=psycarticles-set-c&key=XAP%2F4%2Fxap_4_2_75&mode=pdf

Chan, C. C., Tsui, M. S., Chan, M. Y. C., & Hong, J. H. (2002). Applying the structure of the observed learning outcomes (SOLO) taxonomy on students' learning outcomes: An empirical study. *Assessment and Evaluation in Higher Education, 27*(6), 511–527. Retrieved July 27, 2011, from http://www.tandfonline.com/doi/abs/10.1080/0260293022000020282

Chance, B. (2002). Components of statistical thinking and implications for teaching and assessment. *Journal of Statistics Education, 10*(3). Retrieved August 10, 2008, from http://www.amstat.org/publications/jse/v10n3/chance.html

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–324). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. In B. Chance & A. Rossman (Eds.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics [ICOTS7].* Voorburg, The Netherlands: International Statistical Institute. Retrieved June 5, 2006, from http://www.stat.auckland.ac.nz/~iase/publications/17/7e1_chan.pdf

Chatterji, M. (2004). Evidence on "What works": An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher 33*(9), 3–13. Retrieved August 17, 2010, from http://www.jstor.org/stable/3699819

Cherryholmes, C. H. (1992). Notes on pragmatism and scientific realism. *Educational Researcher*, *21*(6), 13–17. Retrieved August 12, 2011, from http://www.jstor.org/stable/pdfplus/1176502.pdf?acceptTC=true

Chick, H. (1998). Cognition in the formal modes: Research mathematics and the SOLO taxonomy. *Mathematics Education Research Journal, 10*(2), 4–26. Retrieved July 27, 2011, from http://www.springerlink.com/content/90164x9l71256676/fulltext.pdf

Chinnappan, M. (2010). Cognitive load and modelling of an algebra problem. *Mathematics Education Research Journal, 22*(2), 8–23. Retrieved January 12, 2011, from http://www.merga.net.au/publications/counter.php?pub=pub_merj&id=644

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801–823.

Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning, 1*(1), 5–43. Retrieved July 27, 2011, from http://www.tandfonline.com/doi/pdf/10.1207/s15327833mtl0101_1

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375–396). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1–78. Retrieved September 9, 2009, from http://www.jstor.org/stable/3233820

Collis, K. (1975). *A study of concretre and formal operations in school mathematics: A Piagetian viewpoint*. Melbourne: Austrlian Council for Education Research.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed method approaches.* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed method approaches.* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Croninger, R. G., & Valli, L. (2009). Mixing it up about methods. *Educational Researcher, (38)*7, 541–545. Retrieved August 19, 2010, from http://proquest.umi.com/pqdweb?index=11&did=1883613271&srchmode=1&sid=1&fmt=6&vinst=prod&vtype=pqd&rqt=309&vname=pqd&ts=1282188962&clientid=20931

Croninger, R. G., Buese, D., & Larson, J. (2006). *A multi-method look at teaching quality: Insights and quandaries from one study of teaching.* Unpublished manuscript, University of Maryland. Retrieved August 19, 2010, from http://www.education.umd.edu/edci/hqtstudy/pdf%20files/a%20multi-method%20look%20040306%20(2).pdf

Crotty, M. (1998). *The foundations of social research: Meaning and perspective in the research process*. Thousand Oaks, CA: SAGE Publications Inc.

Curcio, F (1987). Comprehension of mathematical relationships expressed in graphs. *Journal of Research in Mathematics Education*, *18* (5), 382–393. Retrieved Sept 5, 2009, from http://www.jstor.org/stable/749086

Darling-Hammond, L., Ancess, J., & Falk, B. (1998). *Authentic assessment in action: Studies of schools and students at work.* New York: Teachers College Press.

delMas, R., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3), Retrieved January 21, 2011, from http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Reasoning about distributions: A collection of research ideas. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy [STRL4]*, Auckland, New Zealand, July 2–7. [CD-ROM]

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 52–82. Retrieved April 10, 2008, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_delMas_Liu.pdf

Denzin, N. K., & Lincoln, Y. S. (2003). *Collecting and interpreting qualitative materials.* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Department of Education, Tasmania. (2003). *Essential Learnings: Introduction to the Outcomes and Standards*. Hobart: Author.

Department of Education, Tasmania. (2008). *The Tasmanian curriculum. Mathematics–numeracy. K – 10 syllabus and support materials.* Retrieved November 2, 2009,from http://resources.education.tas.gov.au/item/edres/29da4473-d8fd-6edd-a1d5-7e6da3384886/1/syl-mnall.pdf

Department of Education, Tasmania. (2010) *Department of Education annual report 2009-10*. Retrieved http://www.education.tas.gov.au/annualreport/09-10/glance0910.pdf

Dever, H. (Project Ed.). (2005). *Fathom 2 Dynamic data software*™ *Fathom Help.* Emeryville, CA: Key College Publishing.

Drijvers, P., Kiernan, C., & Mariotti, M., (with Ainley, J., Andresen, Chan, Y. P., et al.). (2010). Integrating technology into mathematics education: Theoretical perspectives. In M. Artigue & B. R. Hodgson (Gen. Eds.) & C. Hoyles and J-B. Lagrange (Vol. Eds.), *New International Commission on Mathematical Instruction Series: Vol 13. Mathematics Education and Technology – Rethinking the Terrain: The 17th ICMI study* (pp. 89–132). New York, NY: Springer.

Driujvers, P., & Trouche, L. (2008). From artefact to instruments: A theoretical framework behind the orchestra metaphor. In M. K. Heid & G. W. Blume (Eds.), *Research on teaching and learning of mathematics: Cases and perspectives* (Vol. 2, pp. 363–392). Greenwich, CT: Information Age Publishing.

Dunbar, K. N., Fugelsang, J. A., & Stein, C. (2007). Do naive theories ever go away? Using brain and behaviour to understand changes in concepts. In M.C. Lovett & P. Shah, (Eds.), *Thinking with data* (pp. 193–206), New York, NY: Lawrence Erlbaum Associates.

Ercikan, K., & Roth, W-M. (2006). What good is polarising research into qualitative and quantitative? *Educational Researcher, 35*(5), 14–23. Retrieved August 1, 2010, from http://www.jstor.org/stable/3699783

English, L. (2010). Young children's early modelling with data. *Mathematics Education Research Journal, 22*(2), 24–47. Retrieved August 1, 2011, from http://search.informit.com.au/fullText;dn=185470;res=AEIPT

Erickson, T. (2008). *Fifty Fathoms: Statistics Demonstrations for Deeper Understanding*. Emeryville, CA: Key Curriculum Press.

Felicano, G. D., Powers, R. D., & Kearl, B. E. (1963). The presentation of statistical information. *Educational Research and Development*, *11*(3), 32–39. Retrieved September 2, 2009, from http://www.springerlink.com/content/r687u7g22x0w6378/fulltext.pdf

Fielding, A. (1996). Determining adequate sample sizes. *Teaching Statistics, 18*(1), 6–9. Retrieved January 1, 2008, from http://www3.interscience.wiley.com/cgi-bin/fulltext/119953047/pdfstart

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics, 15*(1), 1–24. Retrieved June 12, 2008, from http://www.jstor.org/stable/pdfplus/3482454.pdf

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*, 96–105. Retrieved January 10, 2011, from http://www.iejme.com/032009/p05/fischbein_schnarch_1997.pdf

Fishman, B., Marx, R., Blumenfield, P., Krajcik, J., & Soloway, E. (2004). Creating a framework for research systemic technology innovations. *Journal of the Learning Sciences, 13*(1), 42–76. Retrieved November 12, 2010, from http://pdfserve.informaworld.com/199718_751309879_785041108.pdf

Fitzallen, N., & Brown, N. (2007). Evidence-based research in practice. In P. L. Jeffery (Ed.), *Engaging pedagogies,* Proceedings of the Australian Association for Research in Education [AARE]. Adelaide, SA: AARE. Retrieved August 1, 2011, from www.aare.edu.au/06pap/fit06585.pdf

Flores, A. (2006). Using graphing calculators to redress beliefs in the "law of small numbers." In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth Yearbook* (pp. 291–304). Reston, VA: National Council of Teachers of Mathematics.

Follettie, J. F. (1980). *Bar-graphs-using operations and responses time.* Technical Report. Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development. Eric Document ED250381. Retrieved September 9, 2009, from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED250381&ERICExtSearch_SearchType_0=no&accno=ED250381

Forgasz, H. (2006). Factors that encourage or inhibit computer use for secondary mathematics teaching. *The Journal of Computers in Mathematics and Science Teaching, 25*(1), 77–93. Retrieved December 8, 2010, from http://proquest.umi.com/pqdweb?index=4&did=1005197961&srchmode=3&sid=1&fmt=

Franklin, C. A., & Garfield, J. B. (2006). The GAISE Project: Developing statistics education guidelines for grades pre-K-12 and college courses. In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth Yearbook* (pp. 345–376). Reston, VA: National Council of Teachers of Mathematics.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework*. Retrieved August 20, 2008 from the American Statistics Association Web site http://www.amstat.org/education/gaise/gaiseprek12_intro.pdf

Freebody, P. (2003). *Qualitative research in education. Interaction and practice*. Thousand Oaks, CA: SAGE Publications, Inc.

Freedman, E. G., & Shah, P. (2002). Towards a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer & N. Hari Narayanan (Eds.), *Diagrammatic representation and inference. Second International Conference, Diagrams 2002* (pp. 18–30). Callaway Gardens, GA: Springer-Verlag. Retrieved August 10, 2009, from http://www.springerlink.com/content/dh25j39bavnmpu52/fulltext.pdf

Friel, S. N. (2008). The research frontier: Where technology interacts with the teaching and learning of data analysis and statistics. In M. K. Heid & G. W. Blume (Eds.), *Research on teaching and learning of mathematics: Cases and perspectives* (Vol. 2, pp. 297–332). Greenwich, CT: Information Age Publishing.

Friel, S. N., Curcio, F. R., & Bright, F. R. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124–158. Retrieved from http://www.jstor.org/stable/pdfplus/749671.pdf

Fry, E. (1981). Graphical literacy. *Journal of Reading, 24*(5), 383–389. Retrieved September 13, 2009, from http://ww.jstor.org/stable/40032373

Gal, I. (1998). Assessing statistical knowledge as it relates to students' interpretation of data. In S. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 275–295). Mahwah, NJ: Lawrence Erlbaum.

Gal, I. (2002). Adults' statistical literacy: meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–51. Retrieved September 20, 2009, from http://www.jstor.org/stable/1403713

Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47– 8). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Gal, I. (2005). Towards "probability literacy" for all citizens: Building blocks and instructional dilemmas. In G. A. Jones (Ed.), *Exploring probability in schools: Challenges for teaching and learning* (pp. 39–63). New York, NY: Springer.

Galbraith, P., Stillman, G., Brown, J., & Edwards, I. (2005). *Facilitating mathematical modelling competencies in the middle secondary school.* Retrieved August 10, 2010, from the University of Melbourne, Faculty of Education Web site: http://extranet.edfac.unimelb.edu.au/dsme/ritemaths/general_access/publications/public_papers/2005galstilbroedwconfpaper.pdf

Garfield, J. (1995). How students learn statistics. *International Statistical Review, 63*(1), 25–34. Retrieved July 12, 2006, from http://www.jstor.org/stable/pdfplus/1403775.pdf

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22–38. Retrieved February 13, 2011, from http://www.stat.auckland.ac.nz/~iase/serj/newssep01.pdf

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 397 – 409). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical, Thinking and Learning, 2*(1&2), 99–125. Retrieved August 17, 2011, from http://zsdh.library.sh.cn:8080/FCKeditor/filemanager/upload/jsp/UserImages/1146365172265.pdf

Garfield, J., delMas, R., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In M.C. Lovett & P. Shah (Eds.), *Thinking with data* (pp.117–148). New York, NY: Lawrence Erlbaum Associates.

Glover, D., & Miller, D. (2001). Running with technology: The pedagogic impact of the large-scale introduction of interactive whiteboards in one secondary school. *Journal of Information Technology for Teacher Education, 10*(3), 257–278. Retrieved August 2, 2011, from http://ejscontent.ebsco.com/ContentServer.aspx?target=http%3A%2F%2Fwww%2Etandfonline%2Ecom%2Fdoi%2Fpdf%2F10%2E1080%2F14759390100200115%3F%26userIP%3D131%2E217%2E6%2E7

Goos, M., Dole, S., & Makar, K. (2007). Supporting an investigative approach to teaching secondary school mathematics: A professional development model. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice.* (Proceedings of the thirtieth annual conference of the Mathematics Education Research Group of Australasia [MERGA], Vol. 1, pp. 325–334). Hobart, Tasmania: MERGA. Retrieved January 12, 2011 from http://www.merga.net.au/publications/counter.php?pub=pub_conf&id=399

Graesser, A. C., Swamer, S. S., Baggett, W. B., & Sell, M. A. (1996). New models of deep comprehension. In B. K. Britton & A. C. Graesser (Eds.), *Models of deep comprehension* (pp. 1–32). Mahwah, NJ: Lawrence Erlbaum Associates.

Graham, A. T., & Thomas, M. O. J. (2005). Representational versatility in learning statistics. *The International Journal for Technology in Mathematics Education*, *12*(1), 3-13. Retrieved August 2, 2011, from http://search.proquest.com/docview/203470901/fulltextPDF/130ED51B726400C5799/6?accountid=14245

Green, D. R. (1991). *A longitudinal study of pupil's probability concepts* (Technical Report ME90/91). Loughborough, England: Loughborough University of Technology.

Greene, J. C, Benjamin, L., & Goodyear, L. (2001). The merits of mixing methods in evaluation. *Evaluation 7*(5), 25–44. Retrieved October 7, 2012, from http://evi.sagepub.com/content/7/1/25

Greer, B., & Mukhopadhyay, S. (2005). Teaching and learning the mathematization of uncertainty: Historical, cultural, social and political contexts. In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 297–324). New York: Springer.

Groth, R. E. (2006). Engaging students in authentic data analysis. In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook* (pp. 41–48). Reston, VA: National Council of Teachers of Mathematics.

Guba E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation.* Newbury Park, CA: SAGE Publications Incorporated.

Guin, D., & Trouche, L. (1999). The complex process of converting tools into mathematical instruments: The case of calculators. *International Journal of Computers for Mathematical Learning, 3*, 195–227. Retrieved September 9, 2008, from http://www.springerlink.com/content/m154266063r42355/fulltext.pdf

Harding, D. (1992). Political polls and errors. *Teaching Statistics*, *14*(2), 6.

Hart, L. C., Smith, L. C., Swars, S. L., & Smith, M. E. (2009). An examination of research methods in mathematics education. *Journal of Mixed Methods Research, 3*(1), 26–41. Retrieved April 22, 2010, from http://mmr.sagepub.com/content/3/1/26

Hammerman, J. K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17–41. Retrieved June 12, 2006, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Hammerman_Rubin.pdf

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic enquiry with data: Critical barriers to implementation. *Educational Psychologist, 27*(3), 337–364.

Haspekian, M. (2005). An “Instrumental Approach” to study the integration of computer tools into mathematics teaching: The case of spreadsheets. *International Journal of Computers for Mathematical learning, 10*, 109–141. Retrieved July 29, 2011, from http://www.springerlink.com/content/n618788362q45x1t/fulltext.pdf

Hawkins, A. S., & Kapadia, R. (1984). Children’ conception of probability – a psychological and pedagogical review. *Educational Studies in Mathematics, 15*(4), 349–377. Retrieved January 10, 2011, from http://www.springerlink.com.ezproxy.utas.edu.au/content/0013-1954/15/4/

Hegedus, S. J. (2004). Dynamic representations: A new perspective on instrumental genesis. In M. Bosch (Ed.), *Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education* [CERME 4] (pp. 1031–1040). Retrieved December 22, 2010, from http://ermeweb.free.fr/cerme4/cerme4_wg9.pdf

Hershkowitz, R., Schwarz, B. B., & Dreyfus, T. (2001). Abstraction in context: Epistemic actions. *Journal for Research in Mathematics Education, 32*(2), 195–222. Retrieved December 16, 2010, from http://www.jstor.org/stable/pdfplus/749673.pdf?acceptTC=true

Hershkowitz, R., Hadas, N., Dreyfus, T., & Schwarz, B. (2007). Abstracting processes, from individuals’ constructing of knowledge to a group’s “shared knowledge.” *Mathematics Education Research Journal, 19*(2), 41–68. Retrieved August 12, 2011, from http://www.merga.net.au/documents/MERJ_19_2_Hershkowitz.pdf

Hesterberg, T. (2006). Bootstrapping students’ understanding of statistical concepts. In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook* (pp. 391–416). Reston, VA: National Council of Teachers of Mathematics.

Hollands, J. G., & Spence, I. (1992). Judgments of change and proportion in graphical perception. *Human Factors*, *34*, 313–334.

Hollingsworth, H., Lokan, J., & McCrae, B. (2003). *Teaching mathematics in Australia: Results from the TIMSS 1999 video study*. Melbourne: Australian Council for Educational Research. Retrieved from http://research.acer.edu.au/cgi/viewcontent.cgi?article=1003&context=timss_video

Hosein, A., Aczel, J., Clow, D., & Richardson, J. T. E. (2008). Comparison of black-box, glass-box and open-box software for aiding conceptual understanding. *Proceedings of the 32nd Annual Conference for the Psychology of Mathematics Education [PME 32]*, 17–21 July, 2008, Morelia, Mexico. Retrieved August 26, 2011, from http://oro.open.ac.uk/24542/2/4E449A31.pdf

Howe, K. R. (1992). Getting over the quantitative-qualitative debate. *American Journal of Education, 100*(2), 236–256. Retrieved April 12, 2010 from http://www.jstor.org/stable/pdfplus/1085569.pdf

Howe, K. R. (2009). Positivist dogmas, rhetoric, and the education science question. *Educational Researcher, 38*(6), 428–440. Retrieved August 19, 2010, from http://proquest.umi.com/pqdweb?index=2&did=1858715281&srchmode=1&sid=1&fmt=6&vinst=prod&vtype=pqd&rqt=309&vname=pqd&ts=1282193641&clientid=20931

Hoyles, C., & Noss, R. (2008) Next steps in implementing Kaput's research programme. *Educational Studies in Mathematics, 68*(2), 85–87. http://www.springerlink.com/content/e7q1v48138232250/fulltext.pdf

Hoyles, C., & Noss, R. (2009). The technological mediation of mathematics and its learning. *Human Development, 52*, 129–147. Retrieved http://proquest.umi.com/pqdlink?ver=1&exp=12-14-2015&fmt=7&did=1675591751&rqt=309&cfc=1

Hoyles, C., Noss, R., & Kent, P. (2004). On the integration of digital technologies into mathematical classrooms. *International Journal of Computers for Mathematical Learning, 9,* 309–326. Retrieved October 12, 2009, from http://www.springerlink.com/content/w37333g56mm18275/fulltext.pdf

Ireland, S., & Watson, J. (2009). Building a connection between experimental and theoretical aspects of probability. *International Electronic Journal of Mathematics Education*, *4*(3), 229-260.

Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics in the Middle School, 5*(4), 240–263.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Towards a definition of mixed methods research. *Journal of Mixed Methods Research, 1*, 112–133. Retrieved March 22, 2010 from http://mmr.sagepub.com/cgi/content/abstract/1/2/112

Johnston-Wilder, P., & Pratt, D. (2007). The relationship between local and global perspectives on randomness. In D. Pitta-Pantazi & G. Philipou (Eds.), *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education* (pp. 742–751). Larnaca, Cyprus 2007. Retrieved October 12, 2009, from http://erme.free.fr/CERME5b

Joint Committee for Guides in Metrology. (2008). *Evaluation of measurement data – Guide to the uncertainty in measurement.* (Report by Working Group 1). Retrieved August 12, 2011, from http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

Jones, G. A. (2005). *Exploring probability in school: Challenges for teaching and learning.* New York: Springer

Jones, D. L., & Tarr, J. E. (2007). An examination of the cognitive demand required by probability tasks in middle grade mathematics textbooks. *Statistics Education Research Journal*, *6*(2), 4–27. Retrieved October 13, 2009, from http://www.stat.auckland.ac.nz/~iase/serj/serj6(2)_jones_tarr.pdf

Jones, G. A. Langrall, C. W., & Mooney, E. S. (2007). Research in probability – Responding to classroom realities. In F. K. Lester, Jr. (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 909–955). Information Age Publishing Inc.: Charlotte, NC.

Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development of statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97–117). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kadar, G. D., & Perry, M. (2006). A framework for teaching statistics within the K-12 mathematics curriculum. In B. Chance & A. Rossman (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics [ICOTS7].* Voorburg, The Netherlands: International Statistical Institute. Retrieved January 14, 2011, http://www.stat.auckland.ac.nz/~iase/publications/17/2b3_kade.pdf

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology, 3*, 430–454.

Kapadia, R. (2008). Chance encounters – 20 years later. Fundamental ideas in teaching probability at school level. International Conference of Mathematics Education 11 [ICME11] Monterrey Mexico 2008 Technical Study Group 13 [TSG13]: Research and development in the teaching and learning of probability. Retrieved January 13, 2011, from http://www.stat.auckland.ac.nz/~iase/publications/icme11/icme11_tsg13_24p_kapadia.pdf

Kaput, J. J. (1992). Technology and mathematics education. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 515–556). New York, NY: Macmillan Publishing Company.

Kaput, J. J. (1994). The representational roles of technology in connecting mathematics with authentic experience. In A. J. Bishop (Managing Ed.), *Mathematics Education Library Vol. 13,* R. Biehler, R. W. Scholz, R. SträSSer & B. Winkelmann (Vol. Eds.), *Didatics of Mathematics as a Scientific Discipline.* Dordrecht, The Netherlands: Kluwer Academic Publishers. Retrieved January 11, 2011, from http://bib.tiera.ru/dvd53/biehler%20r.%20(ed.),%20scholz%20r.w.%20(ed.),%20strasser%20r.%20(ed.)%20-%20didactics%20of%20mathematics%20as%20a%20scientific%20discipline(2002)(467).pdf#page=390

Kennewell, S. (2001). Using affordances and constraints to evaluate the use of information and communications technology in teaching and learning. *Technology, Pedagogy and Education, 10*(1&2), 101–116.

Key Curriculum Press. (2005). Fathom® Dynamic Data™ Version 2. [Computer software]. Emeryville, CA: Author.

Key Curriculum Press. (2007). *Exploring statistics with Fathom™ Version 2 Dynamic data software.* Emeryville, CA: Author.

Kidman, G. C., & Nason, R. A. (2000). When a visual representation is not worth a thousand words. *Technology in Mathematics Education [TIME].* Auckland, New Zealand: TIME 2000.

Kieran, C., & Drijvers, P. (2006). The co-emergence of machine techniques, paper-and-pencil techniques, and theoretical reflection: A study of CAS use in secondary school. *International Journal of Computers for Mathematical Learning, 11*(2), 205–263. Retrieved July 27, 2011, from http://www.springerlink.com/content/u7t3580294652u37/fulltext.pdf

Kmietowic, Z. (1994). Sampling errors in political polls. *Teaching Statistics*, *16*(3), 70–74.

Konold, C. (1989). Informal conceptions about probability. *Cognition and Instruction, 6*, 59–98. Retrieved August 3, 2009, from http://www.jstor.org/stable/3233463

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modelling them. *International Journal of Computers for Mathematical Learning*, *12*(3), 217-230. Retrieved August 6, 2009, from http://www.springerlink.com/content/lx2213721548n428/fulltext.pdf

Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics/

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1), 1–37. Retrieved August 3, 2009, from http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1

Konold, C., & Pollatsek, A., (2004). Conceptualizing an average as a stable feature of a noisy process. In D. Ben-Zvi & J, Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 169–200). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Konold, C., Pollatsek, A., Well, A., & Gagon, A. (1996). Students analysing data: Research of critical barriers. In J. Garfield & G. Burril (Eds.), *Research on the role of technology in teaching and learning statistics. Proceedings of the 1996 International Association of Statistics Education round table conference. University of Granada, Spain, July 23–27* (pp. 151–168). Voorburg, The Netherlands: International Statistics Institute. Retrieved January 10, 2011, from www.stat.auckland.ac.nz/~iase/publications/8/13.Konold.pdf

Kosslyn, S. (1994). *Elements of graph design.* New York, NY: Freeman.

Krause, K., Bochner, S., & Duchesne, S. (2003). *Education psychology for learning and teaching.* Southbank, Vic.: Thompson.

Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing.* Thousand Oaks, CA: SAGE Publications Incorporated.

Kvale, S. (2007). Doing Interviews. In (Part 2) *The SAGE qualitative research kit.* London: SAGE Publications Limited.

Laborde, C. (2001). Integration of technology in the design of geometry tasks with Cabri-Geometry. *International Journal of Computers for Mathematical Learning, 6*(3), 283–317. Retrieved July 27, 2011, from http://www.springerlink.com/content/m778453426r2t353/fulltext.pdf

Lagemann, E., & Shulman, L. (Eds.). (1999). *Issues in educational research: Problems and possibilities.* San Francisco: Jossey-Bass Publishers.

Lagrange, J. B. (2005). Curriculum, classroom practices, and tool design in the learning of functions through technology-aided experimental approaches. *International Journal of Computers for Mathematical Learning 10*, 143–189. Retrieved December 10, 2011, from http://www.springerlink.com/content/u07l31845k263813/

Lane-Getaz, S. (2006). What is statistical thinking, and how is it developed? In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook* (pp. 273–290). Reston, VA: National Council of Teachers of Mathematics.

Lavigne, N. C., & Lajoie, S. P. (2007). Statistical reasoning of middle school children engaged in survey enquiry. *Contemporary Educational Psychology, 32*(4), 630–666.

Lehrer, R., Kim, M. J., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modelling variability. *International Journal of Computers for Mathematical Learning, 12*(3), 195–216. Retrieved August 11, 2009, from http://www.springerlink.com/content/126m42t302631133/fulltext.pdf

Lehrer, R., Konold, C., & Kim, M. J. (2006). Connecting data, modelling chance in the middle school. *Paper presented at the 2006 annual meeting of the American Educational Research Association [AERA]*. San Francisco, CA: AERA. Retrieved August 12, 2009, from http://srri.umass.edu/files/lehrer-2006cdm.pdf

Lehrer, R., & Schauble, L. (2002) Distribution: A resource of understanding error and natural variation. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on the Teaching of Statistics (ICOTS6), Cape Town, South Africa.* Voorburg, The Netherlands: International Statistics Institute. Retrieved January 16, 2010, from http://www.stat.auckland.ac.nz/~iase/publications/1/8b3_lehr.pdf

Lesh, R., & Harel, G. (2003). Problem solving, modelling, and local conceptual development. *Mathematical Thinking and Learning, 5*(2&3), 157–189. Retrieved August 17, 2009, from http://www.tandfonline.com.ezproxy.utas.edu.au/doi/pdf/10.1080/10986065.2003.9679998

Lesh, R., Lester, F. K. Jr., & Hjalmarson, M. (2003). A models and modelling perspective on metacognitive functioning in everyday situations where problems solvers develop mathematical constructs. In R. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modelling perspectives on modelling problem solving, learning and teaching* (pp. 383–484). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R., & Zawojewski, J. (2006). Problem solving and modelling. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763–804). Charlotte, NC: Information Age Publishing Inc.

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational research's credibility gaps? *Issues in Education, 5*(2), 177–229.

Lock, R. (2002). Using Fathom to promote interactive explorations of statistical concepts. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on the Teaching of Statistics (ICOTS6), Cape Town, South Africa.* Voorburg, The Netherlands: International Statistics Institute. Retrieved January 13, 2011, from http://www.stat.auckland.ac.nz/~iase/publications/1/7g3_lock.pdf

Lubienski, S. T. (2000). A clash of cultures? Students' experiences ina discussion-intensive seventh-grade mathematics classroom. *The Elementary School Journal, 100*(4), 377–400. Retrieved August 19, 2012 from http://www.jstor.org/stable/pdfplus/1002148.pdf?acceptTC=true

Ma, L. (1999). *Knowing and teaching elementary mathematics. Teachers' understanding of Fundamental Mathematics in China and the United States.* Mahwah, NJ: Erlbaum.

Maxara C., & Biehler, R. (2006). Student' probabilistic simulation and modelling competence after a computer-intensive elementary course in statistics and probability. In B. Chance & A. Rossman (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute. Retrieved October 12, 2008, from http://www.stat.auckland.ac.nz/~iase/publications/17/7c1_maxa.pdf

Maxara, C., & Biehler, R. (2007). Constructing stochastic simulations with a computer tool – Students' competencies and difficulties. In D. Pitta-Pantazi & G. Philipou (Eds.) *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education* (pp. 762–771). Larnaca, Cyprus 2007. Retrieved October 12, 2009, from http://erme.free.fr/cerme5b

Maxara, C., & Biehler, R. (2010). Students' understanding and reasoning about sample size and the law of large numbers after a computer-intensive introductory course of stochastics. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistics Institute. Retrieved October 15, 2010, from http://www.stat.auckland.ac.nz/~iase/publications/icots8/icots8_3c2_maxara.pdf

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics 45*(1/3), 103–129. Retrieved October 10, 2007, from http://www.jstor.org/stable/3483098

McCrone, S. S. (2005). The development of mathematical discussion: An investigation in a fifth-grade classroom. *Mathematical Thinking and Learning 7*(2), 111-133. Retrieved August 15, 2012, from http://www.tandfonline.com/doi/abs/10.1207/s15327833mtl0702_2

MacDonald-Ross, M. (1977). Graphics in texts. *Review of Research in Education 5*, 49–85. Retrieved July 26, 2011, from http://www.jstor.org/stable/pdfplus/1167172.pdf?acceptTC=true

Meira, L. (1998). Making sense of instructional devices: The emergency of transparency in mathematical activity. *Journal for Research in Mathematics Education, 29*(2), 121–142. Retrieved August 12, 2009, from http://www.jstor.org/stable/pdfplus/749895.pdf

Mertens, D. M. (2010). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Metz, K. E. (1997). Dimensions in the assessment of students' understanding and application of chance. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 223–238). Amsterdam, The Netherlands: IOS Press.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary children and undergraduates. *Cognition and Instruction, 16*(3), 285–365. Retrieved June 14, 2010, from http://www.jstor.org/stable/3233647

Meyer, M. (2009). Use of words – language-games in mathematics education. In V. Durand-Guerrier, S. Soury-Lavergne, & F. Arzarello (Eds.), *Proceedings of the Sixth Congress of the European Society for Research in Mathematics Education [CERME6], Lyon, France.* Institut National de Recherche Pédagogique. Retrieved November 12, 2011, from http://www.inrp.fr/publications/edition-electronique/cerme6/wg6-08-meyer.pdf

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20–39. Retrieved July 27, 2011, from http://www.jstor.org/stable/749226

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23–63. Retrieved December, 10, 2009, from http://dx.doi.org/10.1207/S15327833MTL0401_2

Morgan, D. (2007). Paradigms lost and pragmatism regained: Methodological implications for combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), 48–76. Retrieved March 23, 2010, from http://mmr.sagepub.com/cgi/content/abstract/1/1/48

Moss, G., Jewitt, C., Levacic, R., Armstrong, V., Cardini, A., & Castle, F. (2007). *Research Report 816: The Interactive Whiteboards, Pedagogy and Pupil Performance Evaluation: An Evaluation of the Schools Whiteboard Expansion (SWE) Project: London Challenge.* London Institute of Education. Retrieved January 7, 2011, from http://www.pgce.soton.ac.uk/ict/NewPGCE/pdfs%20IWBs/The%20interactive%20whiteboard,%20pedagogy%20and%20pupil%20performance%20evaluation.pdf

National Council of Teachers of Mathematics [NCTM]. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Nickerson, R. S. (1995). Can technology help teach for understanding? In D. Perkins, J . L. Schwartz, M . M. West & M.S. Wiske (Eds.), *Software goes to school – Teaching for understanding with new technologies* (pp. 7–13). New York: Oxford University Press.

Nolan, J., Phillips, G., Watson, J., Denney, C., & Stambulic, S. (with Allen, R., Cahn, R., Ebbage, R., & Iampolsky, E.) (2000). *Math Quest 12: Mathematical Methods VCE Units 3 and 4.* Milton, QLD: John Wiley & Sons.

Nor, N. M., & Idris, N. (2010). Assessing students informal inferential reasoning using SOLO taxonomy based framework. *Procedia – Social and Behavioural Sciences, 2*(2), 4805–4809.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers.* Dordrecht, The Netherlands: Kluwer.

OECD. (2003). *The PISA 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*. Paris: Author.

Olson, D. R. (2004). The triumph of hope over experience in the search for "What works": A response to Slavin. *Educational Researcher, 33*(1), 24–36.

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology*, *8*(5), 375–387. Retrieved April 12, 2010, from http://pdfserve.informaworld.com/546446_751309879_727473132.pdf

Oxford English Dictionary (2nd ed.). (1989). Oxford, UK: Oxford University Press. Retrieved from http://dictionary.oed.com/

Padiotis, I., & Mikropoulos, T. (2010). Using SOLO to evaluate an educational virtual environment in a technology education setting. *Educational Technology & Society, 13*(3), 233–245. Retrieved January 13, 2011, from http://www.ifets.info/download_pdf.php?j_id=48&a_id=1076

Panizzon, D., Callingham, R., Wright, T., & Pegg, J. (2007). Shifting sands: Using SOLO to promote assessment for learning with secondary mathematics and science teachers. In P. L. Jeffery (Ed.) *Engaging Pedagogies,* Proceedings of the Australian Association for Research in Education [AARE]. Adelaide, SA: AARE.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning education and human activity. *The Journal of the Learning Sciences, 13*(3), 423–451. Retrieved May 14, 2010 from http://pdfserve.informaworld.com/439446_751309879_785041710.pdf

Peebles, D., & Cheng, P. C-H. (2003). Modelling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors, 45*(1), 28-45. [Electronic version] Retrieved September 11, 2009, from http://www.cogs.susx.ac.uk/users/peterch/papers/hfj2003.pdf

Pegg, J. (2003). Assessment in mathematics. In J. M. Rogers (Ed.), *Mathematical Cognition,* 227–259. Greenwich, CA: Information Age Publishing

Perrenet, J., & Adan, I. (2010). The academic merits of modelling in higher mathematics education: A case study. *Mathematics Education Research Journal 22*(2), 121–140. Retrieved January 13, 2011, from http://www.springerlink.com/content/4p48g72231854r06/fulltext.pdf

Pfannkuch, M. (2005). Characterizing Year 11 students' evaluation of statistical process. *Statistics Education Research Journal 4*(2), 5–22. Retrieved January 13, 2011, from http://www.stat.auckland.ac.nz/~iase/serj/serj4(2)_pfannkuch.pdf

Pfannkuch, M. (2008). Building sampling concepts for statistical interference: A case study. *Eleventh International Conference on Mathematical Education*, Monterrey, Mexico, July, 2008. Retrieved April, 12, 2010, from http://tsg.icme11.org/document/get/476

Pfannkuck, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–16). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Phillips, D. C., & Barbules, N. C. (2000). *Postpositivism and educational research*. Lanham, MD: Rowman & Littlefield.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: Norton.

Powell, K. C. (2004). Developmental psychology of adolescent girls: Conflicts and identity issues. *Education 125*(1), 77–87. Retrieved August 2, 2012, from http://search.proquest.com/docview/196438040/fulltextPDF/138A2138CDF56DE15FC/2?accountid=14245

Pratt, D. (2000). Making sense of the total of two dice. *Journal for Research in Mathematics Education, 31*(5), 602–625. Retrieved January 11, 2011, from http://www.jstor.org/stable/pdfplus/749889.pdf

Pratt, D., & Noss, R. (2002). The microevolution of mathematical knowledge: The case of randomness. *The Journal of the Learning Sciences*, *11*(4), 453–488. Retrieved August 25, 2009, from http://www.jstor.org/stable/1466746

Pratt, D., & Noss, R. (2010). Designing for mathematical abstraction. *International Journal of Computers for Mathematical Learning, 15*, 81–97. Retrieved December 15, 2010, from http://www.springerlink.com/content/085p7r20627r0867/fulltext.pdf

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology*, *14*(1), 36–49. Retrieved September 5, 2009, from http://www.apa.org/journals/features/xap14136.pdf

Reading, C. (2004). Students' description of variation while working with weather data. *Statistics Education Research Journal, 3*(2), 86–105. Retrieved January 13, 2011, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Reading.pdf

Reeves, T. C. (2006) Design research from a technology perspective. In J. Van den Akker, K. Gravemeijer, S. McKenny, & N. Nieven (Eds.), *Educational Design Research* (pp. 52-66). New York, NY: Routledge.

Richardson, L. (2003). Writing. A method of inquiry. In N. K. Denzin & Y. S. Lincoln (Eds.). *Collecting and interpreting qualitative materials.* (2nd ed. pp.). Thousand Oaks, CA: SAGE Publications, Inc.

Ritchhart, R. (2001). From IQ to IC: A dispositional view of intelligence. *Roeper Review, 23*(3), 143–150.

Ritchhart, R. (2002). *Intellectual character: What it is, why it matters, and how to get it.* San Francisco, CA: Jossey-Bass.

Rossman, A., Chance, B., & Medina, E. (2006). Some important comparisons between statistics and mathematics, and why teachers should care. In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook* (pp. 323 – 324). Reston, VA: National Council of Teachers of Mathematics.

Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5–19. Retrieved August 10, 2009, from www.stat.auckland.ac.nz/~iase/serj/serj7(2)_rossman.pdf

Roth, W-M., & McGinn, M. K. (1998). Inscriptions: Towards a theory of representing as social practice. *Review of Educational Research, 68*(1), 35–59. Retrieved Oct. 3, 2009, from http://www.jstor.org/stable/pdfplus/1170689.pdf

Rowe, K. (2007). The imperative of evidence-based instruction leadership: Building capacity within professional learning communities via a focus on effective teaching practice. *Student Learning Processes.* Retrieved June 10, 2010, from http://research.acer.edu.au/learning_processes/2

Rubin, H. J., & Hammerman, J. K. (2006). Understanding data through new software representations. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook* (pp. 241–246). Reston, VA: National Council of Teachers of Mathematics.

Rubin, H. J., & Rubin, I. S. (2005). *Qualitative interviewing: The art of hearing data.* Thousand Oaks, CA: SAGE Publications, Inc.

Ruthven, K. (2008). The interpretative flexibility, instrumental evolution and institutional adoption of mathematical software in educational practice: The example of computer algebra and dynamic geometry. *Journal of Educational Computing Research 39*(4), 379–394. Retrieved January 14, 2010, from http://www.educ.cam.ac.uk/people/staff/ruthven/ruthvenjecrpreprint.pdf

Ruthven, K., & Hennessy, S. (2002). A practitioner model of computer-based tools and resources to support mathematics teaching and learning. *Educational Studies in Mathematics, 49*(1), 47–88. Retrieved January 6, 2010, from http://www.jstor.org/stable/i277403

Scheaffer, R. L. (2006). Statistics and mathematics: On making a happy marriage. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook.* (pp. 309–322). Reston, VA: National Council of Teachers of Mathematics.

Seeto, D., & Herrington, J. (2006). Design-based research and the learning designer. In L. Markauskaite, P. Goodyear, & P. Reimann (Eds.), *Proceedings of the twenty-third annual Australian society for computers in tertiary education conference: Who's learning? Whose technology?* (pp. 741–745). Sydney: Sydney University Press. Retrieved October 12, 2009, from http://www.ascilite.org.au/conferences/sydney06/proceeding/pdf_papers/p177.pdf

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14*(1), 47–69. Retrieved July 25, 2009, from http://www.springerlink.com/content/v2581778612k5432/fulltext.pdf

Shaughnessy, J. M. (2003). Research on students' understanding of probability. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 216–226). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M. (2006). Research on students' understanding of some big concepts in statistics. In G. F. Burrill & P. C. Elliott, (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 77–98). Reston, VA: NCTM.

Shaughnessy, J. M. (2007). Research in probability: responding to classroom realities. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp.936–999). Charlotte, NC: Information Age Publishing Inc.

Shaughnessy, J. M., & Chance, B. (2005). *Statistical questions from the classroom.* Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on the Teaching of Statistics (ICOTS6), Cape Town, South Africa.* Voorburg, The Netherlands: International Statistics Institute. Retrieved December, 30, 2010 from http://www.stat.auckland.ac.nz/~iase/publications/1/6a6_shau.pdf

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education. National Research Council Committee on Scientific Principles for Education Research.* Washington, DC: National Academy Press.

Sherin, M. G. (2002). A balancing act: Developing a discourse community in a mathematics classroom. *Journal of Mathematics Teacher Education, 5,* 205–233. Retrieved November 9, 2008, from http//www.springerlink.com/content/q1x43544m7386vq6/fulltext.pdf

Silver, E. A. & Smith, M. S. (1996). Building discourse communities in mathematical classrooms: A worthwhile but challenging journey. In P. C. Elliott & M. J. Kenney, (Eds.), *Communication in mathematics, K-12 and beyond: 1996 Yearbook* (pp. 20–28). Reston, VA: NCTM.

Simon, J. L. (1997). *Resampling: The New Statistics*. Published on-line. Retrieved January 14, 2008, from http://www.resample.com/content/text/28-chap-24.pdf

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association, 82*(398), 454–465. Retrieved September 9, 2009, from http://www.jstor.org/stable/pdfplus/2289447.pdf?acceptTC=true

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15–21. Retrieved August 17, 2010, from http://www.jstor.org/stable/3594400

Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational enquirers. *Educational Researcher, 15*(1), 4–13. Retrieved April 12, 2010, from http://www.jstor.org/stable/pdfplus/1174482.pdf

Smith, M. H. (2004). A sample/population size activity: Is it the sample size of the sample as a fraction of the population that matters? *Journal of Statistics Education, 12*(2). Retrieved January 29, 2008, from http://www.amstat.org/publications/jse/v12n2/smith.html

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical interference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98–113. Retrieved December 20, 2010, from https://lirias.kuleuven.be/bitstream/123456789/136347/1/castrosotos.pdf

Stacey, K. (2003). The need to increase attention to mathematical reasoning. In H. Hollingsworth, J. Lokan, & B. McCrae (Eds.), *Teaching mathematics in Australia: Results from the TIMSS 1999 Video Study* (pp. 119–122). Melbourne: Australian Council for Educational Research. Retrieved http://research.acer.edu.au/cgi/viewcontent.cgi?article=1003&context=timss_video

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: SAGE Publications, Inc.

Star, S. L. (1989). The structure of ill-structured solutions: Boundary objects and heterogeneous distributed problem-solving. In L. Glasser & M. N. Huhns (Eds.), *Distributed Artificial Intelligence, Vol. 2*. London: Pitman / San Mateo, CA: Morgan Kaufmann.

Stark, P. B. (2010.). *Glossary of statistical terms.* Retrieved October 10, 2009, from http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm

Steen, L. A. (1999). Numeracy: The new literacy for a data-drenched society. *Educational Leadership, 57*(2), 8–13.

Stillman, G., Brown, J., Galbraith, P. (2010). Researching applications and mathematical modelling. *Mathematics Education Research Journal, 22*(2), 1–7. Retrieved January 12, 2011, from http://www.merga.net.au/documents/merj22(2)-editorial.pdf

Stillman, G., Galbraith, P., Brown, J., & Edwards, I. (2007). A framework for success in implementing mathematical modelling in the secondary classroom. In J. Watson & K. Beswick (Eds.), *Mathematics: essential research, essential practice. Proceedings of the thirtieth annual conference of the Mathematics Education Research Group of Australasia [MERGA]* (Vol. 2, pp. 688–697). Hobart, Tasmania: MERGA.

Stohl, H. S., & Hollebrands, K. F. (2006). Students' use of technological features while solving a mathematics problem. *Journal of Mathematical Behaviour 25*, 252–226. Retrieved July 27, 2011, from http://www.sciencedirect.com/science?_ob=PublicationURL&_tockey=%23TOC%236566%232006%23999749996%23638191%23FLA%23&_cdi=6566&_pubType=J&_auth=y&_acct=C000052220&_version=1&_urlVersion=0&_userid=1526876&md5=6debaf1bf72abc880ad590297fe05897

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference using probability simulation tools. *Journal of Mathematical Behaviour, 21*, 319–337.

Stohl, H., Rider, R., & Tarr, J. (2004). *Making connections between empirical and theoretical probability: Students' generation and analysis of data in a technological environment*. Retrieved August 20, 2009, from http://www.probexplorer.com/articles/leeridertarrconnecte&t.pdf

Tabach, M., Arcavi, A., & Hershkowitz, R. (2008). Transitions among different symbolic generalizations by algebra beginners in a computer intensive environment. *Educational Studies in Mathematics, 69*(3) 53–71. Retrieved August 12, 2011 from http://www.springerlink.com/content/n767035u57121201/fulltext.pdf

Tarr, J. E., Lee, H. S., & Rider, R. L. (2006). When data and chance collide: Drawing inferences from empirical data. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance. Sixty-eighth yearbook* (pp. 139–150). Reston, VA: National Council of Teachers of Mathematics.

Taylor, J. R. (1982). *An introduction to error analysis: The study of uncertainties in physical measurement*. Mill Valley, CA: University Science Books.

Teachers Registration Board of Tasmania. (2006). *Code of Professional Ethics for the Teaching Profession in Tasmania.* Retrieved August 5, 2010, from http://www.trb.tas.gov.au/TRB%20Code%20of%20Professional%20Ethics.pdf

Teachers Registration Board of Tasmania. (2007). *Tasmanian Professional Teaching Standards Framework.* Retrieved August 5, 2010, from http://www.trb.tas.gov.au/Final%20Standards%20July12%2007.pdf

Teddlie, C, & Tashakkori, A. (2010). Overview of contemporary issues in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social & Behavioral Research* (2nd ed., pp. 1–41). Thousand Oaks, CA: SAGE Publications Incorporated.

Theberge, C. L. (1994). Small-group vs. whole-class discussion: Gaining the floor in science lessons. Paper presented at the annual meeting of the American Educational Research Association (AERA), New Orleans, LA. (ERIC Document Reproduction Service No. ED373983). Retrieved August 21, 2012, from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED373983

Thomas, M., & Chinnappan, M. (2008). Teaching and learning with technology: Realising the potential. In H. Forgasz, A. Barkatas, A. Bishop, B. Clarke, S. Keast, W. T. Seah, et al.(Eds.) *Research in Mathematics Education in Australasia 2004-2007* (pp. 165–194). Rotterdam, The Netherlands: SENSE Publishers.

Trouche, L. (2004). Managing the complexity of human/machine interactions in computerized learning environments: Guiding students' command process through instrumental orchestrations. *International Journal of Computers for Mathematical Learning, 9*(3), 281–307. Retrieved July 27, 2011, from http://www.springerlink.com/content/n021r76t0l21jv16/fulltext.pdf

Trouche, L. (2005). Instrumental genesis, individual and social aspects. In D. Guin, K. Ruthven, & L. Trouche (Eds.), *The Didactical Challenge of Symbolic Calculators – Turning a Computational Device into a Mathematical Instrument Vol. 36. Mathematics Education Library* (pp. 197–230). New York, NY: Springer,

Trickett, S. B., & Trafton, J. G. (2006). Towards a comprehensive model of graph comprehension: Making the case for spatial cognition. In D. Barker-Plummer, R. Cox, & N. Swoboda. (Eds.), *Diagrammatic representation and inference. 4th International Conference, Diagrams 2006. Proceedings* (pp. 286-300). Stanford, CA: Springer. [Electronic version] Retrieved Sept 15, 2009, from http://www.springerlink.com/content/y2g3312368321l5w/fulltext.pdf

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105–110. Retrieved December 15, 2010, from http://www.stats.org.uk/statistical-inference/tverskykahneman1971.pdf

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, *57*(2), 74–79.

Vale, C. M., & Leder, G. C. (2004). Student views of computer-based mathematics in the middle years: Does gender make a difference? *Educational Studies in Mathematics, 56*(2/3), 287–312. Retrieved August 2, 2012, from http://www.jstor.org/stable/4150285

Verillon, P., & Rabardel, P. (1995). Cognition and artifacts: A contribution to the study of thought in relation to instrumented activity. *European Journal of Psychology of Education, 10*, 77–103.

Vessey, I. (1991). Cognitive fit: A theory based analysis of the graphs versus tables literature. *Decision Sciences*, *22*, 219–240.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process.* M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.). Harvard, MA: Harvard University Press.

Walpole, R. E., & Myers, R. H. (1978). *Probability and statistics for engineers and scientists* (2nd ed.). New York: Macmillan Publishing Co., Inc.

Walshaw, M. (2007). Responding to calls for greater accountability. *Mathematics Education Research Journal, 19*(1), 1–2.

Walshaw, M., & Anthony, G. (2008). The teacher's role in classroom discourse: A review of recent research into mathematics classrooms. *Review of Educational Research, 78*(3), 516–551. Retrieved October 15, 2010, from http://proquest.umi.com/pqdweb?index=11&did=1580752911&srchmode=3&sid=1&fmt=3&vinst=prod&vtype=pqd&rqt=309&vname=pqd&ts=1291174291&clientid=20931&aid=3

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research & Development, 53*(4), 5–23. Retrieved from August 7, 2010 from http://lopezlearning.net/files/19511441FenWangArticle-2.pdf

Watson, J. (1997). Assessing statistical literacy through the use of media surveys. In I. Gal & J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121) Amsterdam, The Netherlands: International Statistics Institute.

Watson, J. (2005). Developing an awareness of distribution. In K. Makar (Ed.), *Reasoning about distributions: A collection of research ideas. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy, Auckland, New Zealand,* July2–7. [CD-ROM]

Watson, J. M. (2002). When 2 + 2 ≠ 4 and 6 + 6 ≠ 12 in data and chance. *New England Mathematics Journal, 34*(2), 56–58

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals.* Mahwah, NJ: Lawrence Erlbaum Associates.

Watson, J., Beswick, K., Brown, N., Callingham, R., Muir, T. & Wright, S. (2011). *Digging into Australian data with TinkerPlots.* Melbourne, Australia: Objective Learning Materials.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3–46. Retrieved January 13, 2011, from http://www.stat.auckland.ac.nz/~iase/publications/rt04/4.1_watson&callingham.pdf

Watson, J., & Fitzallen, N. (2010). *The development of graph understanding in the mathematics curriculum.* Report for the NSW Department of Education and Training. Retrieved August 12, 2011, from http://www.curriculumsupport.education.nsw.gov.au/primary/mathematics/assets/pdf/dev_graph_undstdmaths.pdf

Watson, J., & Moritz, J. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, *31*(1), 44–70.

Watson, J.M., & Kelly, B. A. (2009). Development of student understanding of outcomes involving two or more dice. *International Journal of Science and Mathematics Education*, *7*(1), 25–54. Retrieved October 10, 2010, from http://www.springerlink.com.ezproxy.utas.edu.au/content/t52q137454538350/fulltext.pdf

Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical interference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168. Retrieved August 12, 2006, from http://www.jstor.org/stable/pdfplus/3483313.pdf?acceptTC=true

Watson, J. M., & Moritz, J. B., (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgment. *Journal for Research in Mathematics Education, 34*(4), 270–340. Retrieved August 31, 2009, from http://www.jstor.org/stable/pdfplus/30034785.pdf

White, T. (2007). Debugging an artifact, instrumenting a bug: Dialectics of instrumentation and design in technology-rich learning environments. *International Journal of Computers for Mathematical Learning, 13*, 1–26. Retrieved July 27, 2011, from http://www.springerlink.com/content/9710210t60v3x543/fulltext.pdf

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–265.

Wilkins, J. L. (2000). Preparing for the 21st century: The status of quantitative literacy in the United States. *School Science and Mathematics*, *100*(8), 405–418.

Wood, M. (2005). The role of simulation approaches in statistics. *Journal of Statistics Education, 13*(3). Retrieved January 29, 2008, from http://www.amstat.org/publications/jse/v13n3/wood.html

Yoon, C., Dreyfus, T., & Thomas, M. O. J., (2010). How high is the tramping track? Mathematising and applying in a calculus model-eliciting activity. *Mathematics Education Research Journal, 22*(2), 141–151. Retrieved August 1, 2011, from http://search.informit.com.au/fullText;dn=185470;res=AEIPT

Zbiek, R., & Conner, A. (2006). Beyond motivation: Exploring mathematical modelling as a context for deepening students' understandings of curricular mathematics. *Educational Studies in Mathematics, 63*(1), 89–112. Retrieved September 10, 2010, from http://www.springerlink.com/content/k25t2r61q1g7m278/fulltext.pdf

Zieffler, A., & Garfield, J. (2007). Studying the role of simulation in developing students' statistical reasoning. Retrieved August 12, 2008, from http://www.stat.auckland.ac.nz/~iase/publications/isi56/IPM40_Zieffler.pdfrieved

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education*, *16*(2). Retrieved July 14, 2009, from www.amstat.org/publications/jse/v16n2/zieffler.html

Zimmermann, G. M., & Jones, G. A. (2002). Probability simulation: What meaning does it have for high school students? *Canadian Journal of Science, Mathematics and Technology Education, 2*(2), 221–236.

Acronyms and initialisations

Acronym or initialisation	Full title	Page reference
ABS	Australian Bureau of Statistics	ii
ACARA	Australian Curriculum, Assessment and Reporting Authority	67
APAI	Australian Postgraduate Award Industry	ii
ARC	Australian Research Council	ii
CA	California, USA	ii
EA	Extended Abstract See SOLO, Structure of Observed Learning Outcome	74
GAISE	Guidelines for Assessment and Instruction in Statistics Education	18
GICS	Global, Individual, Measures of Centre and Measures of Spread	89
HERCS	Human Research Ethics Committee (Tasmania) Network	ii
ICSEA	Index of Community Socio-Educational Advantage	87
IWB	Interactive Whiteboard	57
KCP	Key Curriculum Press – owner and software developer of Fathom	vi
M	Multistructural See SOLO, Structure of Observed Learning Outcome	74
M_TR	Transitional response between Multistructural and Relational. See SOLO, Structure of Observed Learning	76
MS-Excel	Microsoft Excel	251
MS-PowerPoint	Microsoft PowerPoint	82
NCTM	National Council of Teachers of Mathematics	19
NRC	National Research Council	12
OECD	Organisation of Economic and Cultural Development	15
OED	Oxford English Dictionary	
PS	Prestructural See SOLO, Structure of Observed Learning Outcome	74
R	Relational See SOLO, Structure of Observed Learning Outcome	74
SOLO	Structure of Observed Learning Outcome	71
SRA	Scientific Research Approach	10
U	Unistructural See SOLO, Structure of Observed Learning Outcome	74

Glossary

Term	**Section first used**	Definition	**Definition reference**
Attribute, Fathom	3.2.10, p. 92	A descriptor of a case, for example, the height of a person is an attribute	Dever, 2005
Boundary object	2.4.4, p. 24	Provides shared common ground around which mathematical meaning may be constructed e.g., Fathom simulation, a graphic [See also situated abstraction]	Hoyles, Noss & Kent, 2004; Star, 1989
Case, Fathom	3.2.10, p. 92	An individual record in a collection , for example, a set of measurements taken at a given time	Dever, 2005
Collection, Fathom	3.2.10 p. 92	The container for data in Fathom, alternatively a Fathom data set	Dever, 2005
Confirmability	2.2.6, p. 12	Whether the survey instrument allows inferences to be drawn	Mertens, 2010
Credibility	2.2.6, p. 12	Whether the study investigated what was intended to be investigated	Gubq & Lincoln, 1989
Dependability	2.2.6, p.12	The consistency of responses to the research protocol	Gubq & Lincoln, 1989
Deterministic	2.3.3, p. 16	Where each result must have a explainable cause. In contrast probabilistic, where unexplainable factors influence the result. Equated with mathematical and statistical thinking respectively.	Schaeffer, 2006
Document		A Fathom file	Dever, 2005
Drag-and-drop	3.7.3, p. 128	A software operation where an icon or cases are moved across a computer workspace	Author
Equiprobability bias	2.4.7 p. 30	A misconception where each die is fair, so the outcome of each individual die was equally likely so the outcome of the sum of the two dice is equally likely.	LeCoutre, 1992, cited in Pratt, 200

Glossary (cont.)

Error	2.4.9, p. 36	The uncertainty associated with any measurement. In this context should not be confused with mistake or blunder.	Taylor, 1982
Estimate	2.4.9, p. 36	Any measurement is more accurately an estimate because the true underlying value cannot ever be known with perfect certainty. In common language an estimate is a judgement by eye, and not a formal measurement. The various definitions converge on the word approximation. See also error.	Taylor, 1982
Fairness	2.4.7, p. 31	Articulation of randomness. Often cued using the physical characteristics of the physical simulator of the die or coin	Pratt & Noss, 2002
Fairness measure	3.2.9, p. 90	Statistic developed for this study. Calculated as the sum of the difference between observed and expected frequency	Author
Fathom workspace	3.3.3.5 p. 108	A Fathom document window as displayed on a computer screen. Fathom is open and operating, but no Fathom files or documents are in use	Author
Generalisability	2.2.6, p.12	Refers to the transferability to a new setting or people	Kvale, 2006
Icon, iconic	2.4.7 p. 30	A symbolic representation. The symbol is not arbitrary, but bears some resemblance to what it stands for, e.g., a virtual dice. If the icon behaves as the same then the icon is also an analogue.	OED
Instrumental genesis	2.5.6.1 p. 58	The process by which skills and knowledge are applied to a bare tool to produce an effective instrument.	Drijvers & Trouche, 2008
irregularity	2.4.7, p. 31	Articulation of randomness. No discernible regular pattern observed	Pratt & Noss, 2002
Measure (noun)	3.2.9, p. 90	A statistic	Fathom
Measurement	2.4.9, p.36	More strictly an approximation.	Taylor, 1982

Glossary (cont.)

Modelling	2.4.11, 37	Within this study refers to formal algebraic modelling.	Author
Natural language	2.4.9, 36	Informal everyday speech in extended use	OED
Objectivity	2.2.6 p. 12	The conclusions are evidence-based and explicitly linked to the data	Shavelson & Towne, 2002
Physical simulation	2.4.7 p. 30	A physical object is used to simulate a mathematically random process, e.g., a physical die used to simulate a random process of six equally likely outcomes [See also virtual simulation]	Author
Practical importance	2.4.9 p. 36	The practical significance of the accuracy of measurement [see also statistical significance]	Author
Probabilistic	2.3.3 p. 16	Subject to or involving chance variations or uncertainties. A probabilistic decision is one made under uncertainty[see also deterministic]	Schaeffer, 2006
Reliability	2.2.6 p. 12	The consistency of responses to the research protocol	Kvale, 1996
Scaffolding	2.5.6.2 p. 64	The assistance of an expert, adult, or teacher that provides the appropriate level of support for students to extend their knowledge [See also webbing]	Pea, 2004
Situated abstraction	2.5.6.2 p. 63	Construct mathematical ideas by drawing on the webbing of a particular situation.	Noss & Hoyles, 1996
Statistic	3.2.9, p. 90	One number that represents a more complex number	Oxford concise mathematics
Summary, Fathom	3.3.3.5 p. 108	A table of summary statistics. Also known as a summary table.	Dever, 2005
Table, Fathom	3.3.9	A data collection presented in a tabular form, i.e., in a manner similar to an Excel spreadsheet. Also known as a Case table	Dever, 2005
Taskbar / Object shelf	Appendix A.6	The location of key icons. Used to be consistent with MS products, but known formally in Fathom as the Object shelf.	Dever, 2005
Transferability	2.2.6, p. 12	See generalisability	

Glossary (cont.)

Triangulation	2.2.6, p. 12	The use of a variety of research methods to enhance the credibility of the research	
Unpredictability	2.4.7 p. 12	Articulation of randomness. Subsequent outcome is not predictable	Pratt & Noss, 2002
Unsteerability	2.4.7 p. 12	Articulation of randomness. No known agent determines the outcome	Pratt & Noss, 2002
Validity	2.2.6 p. 12	The accuracy and credibility of the research findings	Cresswell, 2003
Virtual simulation	2.4.7, p. 30	A virtual simulation of, in this study, of a random process. As distinct from a physical simulation using a die.	Author
Webbing	2.5.6.2, p.63	A structure that supports learning	Noss & Hoyles, 1996
Worksheet		Hard-copy of instructions provided to students participating in the study	Author
Workspace	3.3.3.5 p. 108	See Fathom workspace	Author

Mathematical symbols

Symbol	**Definition**
X	Normally distributed random variable
μ	mean
σ^2	variance
$\hat{p}$	sample proportion of success
$\hat{q}$	sample proportion of failure
Z	Transformed normal random variable
σ	standard deviation
p	probability of success
q	probability of failure
n	sample size
e	margin of error

Super and subscripts	**Definition**
i	individual
^	sample, as distinct from population [circumflex accent]