CHANGES WITH AGE

IN STUDENTS' MISCONCEPTIONS

OF DECIMAL NUMBERS

VICKI STEINLE

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

March, 2004

Department of Science and Mathematics Education
The University of Melbourne

ABSTRACT

This thesis reports on a longitudinal study of students' understanding of decimal notation. Over 3000 students, from a volunteer sample of 12 schools in Victoria, Australia, completed nearly 10000 tests over a 4-year period. The number of tests completed by individual students varied from 1 to 7 and the average inter-test time was 8 months. The diagnostic test used in this study, (Decimal Comparison Test), was created by extending and refining tests in the literature to identify students with one of 12 misconceptions about decimal notation.

Particular longitudinal measures and definitions of the prevalence of misconceptions were adapted from the medical literature. These measures were further refined to overcome the effect of repeated testing (which resulted in a 10% improvement) as well as various sampling issues. Analysis was conducted at both the coarse level (4 behaviours) and fine level (12 ways of thinking).

Improved estimates of the prevalence of expertise as well as for the various misconceptions are provided. Only 30% of Grade 6 students and 70% of Grade 10 students demonstrate expertise on this test and about 25% of students between Grades 7 and 10 completed tests by choosing the decimal with the *fewest* digits as the *largest* number, a behaviour which results from several different ways of thinking. Despite its high prevalence, this particular behaviour is not well known amongst teachers.

Three phenomena were investigated: persistence, hierarchy and regression. The misconceptions which are most persistent are those that involve the treatment of the decimal portion of a number as a whole number. A hierarchy of the misconceptions was determined by considering the relative rate to expertise on the next test: the hierarchy is different for primary and secondary students. About 20% of students were involved in regression, that is, they completed one test as an expert, but were unable to do so on a later test. This analysis provides additional evidence that many students are receiving teaching that *covers over* rather than *overcomes* their misconceptions. For example, some students appear to be following algorithms for comparing decimal numbers (such as rounding to two decimal places), but revert to a latent misconception when their incomplete algorithm fails. Furthermore, support is provided for the hypothesis that some misconceptions are due to the interference of new teaching.

DECLARATION

This thesis contains no material that has been accepted for any other degree in any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where reference is given in the text. The thesis is less than 100 000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signature:

Date:

ACKNOWLEDGEMENTS

I would like to thank Professor Kaye Stacey for her guidance and encouragement over the years. Her enthusiasm sustained me when the task seemed rather daunting. I would like to also express my thanks to my husband Peter who not only assisted with reading drafts of this thesis but also spent many hours looking after two small children so that I could commence this research many years ago. To my two "thesis orphans" Joanna and Kathleen: I would like to say thank you for your patience and hope that we still have plenty of time left to play. Thanks to Mike Rumsewicz for reading and commenting on a draft, and to my parents Doreen and Colvyn Schofield for encouraging me to strive for high ideals.

I have been most fortunate to have the opportunity to engage in fruitful conversations with my colleagues in the Department of Science and Mathematics Education at the University of Melbourne. In particular, I would like to thank John Dowsey for helpful statistical discussions.

Publications by candidate on matters relevant to thesis:

Asp, G., Chambers, D., Scott, N., Stacey, K., & Steinle, V. (1997). Using Multimedia for the Teaching of Decimal Notation, *Annual conference of the Mathematics Association of Victoria*.

Stacey, K., Helme, S., & Steinle, V. (2001). Confusions between decimals, fractions and negative numbers: A consequence of the mirror as a conceptual metaphor in three different ways. In M. v. d. Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 217-224). Utrecht: PME.

Stacey, K., Helme, S., Steinle, V., Baturo, A., Irwin, K., & Bana, J. (2001). Preservice Teachers' Knowledge of Difficulties in Decimal Numeration. *Journal of Mathematics Teacher Education, 4*(3), 205-225.

Stacey, K., Sonenberg, E., Nicholson, A., Boneh, T. & Steinle, V. (2003). A teaching model exploiting cognitive conflict driven by a Bayesian network. In P. Brusilovsky, A. T. Corbett, and F. De Rosis (Eds.) *Lecture Notes in Artificial Intelligence (Ninth International Conference on User Modeling UM2003) 2702/2003*, 352–362. Springer-Verlag, Heidelberg.

Stacey, K., & Steinle, V. (1998). Refining the Classification of Students' Interpretations of Decimal Notation. *Hiroshima Journal of Mathematics Education, 6*, 49-69.

Stacey, K., & Steinle, V. (1999a). A longitudinal study of children's thinking about decimals: A preliminary analysis. In O. Zaslavsky (Ed.), *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 233-240). Haifa: PME.

Stacey, K., & Steinle, V. (1999b). Understanding decimals: The path to expertise. In J. M. Truran & K. M. Truran (Eds.), *Making the difference. Proceedings of the 22nd annual conference of the Mathematics Education Research Group of Australasia* (pp. 446-453). Adelaide: MERGA.

Steinle, V., & Stacey, K. (1998a). The incidence of misconceptions of decimal notation amongst students in Grades 5 to 10. In C. Kanes, M. Goos, & E. Warren (Eds.), *Teaching Mathematics in New Times. Proceedings of the 21st annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 548-555). Gold Coast, Australia: MERGA.

Steinle, V., & Stacey, K. (1998b). Students and decimal notation: Do they see what we see? In J. Gough & J. Mousley (Eds.), *35th annual conference of the Mathematics Association of Victoria* (Vol. 1, pp. 415-422). Melbourne: MAV.

Steinle, V., & Stacey, K. (2001). Visible and invisible zeros: Sources of confusion in decimal notation. In J. Bobis, B. Perry & M. Mitchelmore (Eds.), *Numeracy and Beyond. Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 434-441). Sydney: MERGA.

Steinle, V., & Stacey, K. (2002). Further evidence of conceptual difficulties with decimal notation. In B. Barton, K. Irwin, M. Pfannkuch & M. Thomas (Eds.), *Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 633-640). Auckland: MERGA.

Steinle, V., & Stacey, K. (2003a). Exploring the right, probing questions to uncover decimal misconceptions. In L. Bragg, C. Campbell, G. Herbert & J. Mousley (Eds.), *Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 634-641). Geelong: MERGA..

Steinle, V., & Stacey, K. (2003b). Grade-related trends in the prevalence and persistence of decimal misconceptions, *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 259-266). Hawai'i: PME.

Steinle, V., & Stacey, K. (2004a). A longitudinal study of students understanding of decimal notation: an overview and refined results. In I. Putt, R. Faragher & M. McLean (Eds.), *Proceedings of the 27th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 541-48). Townsville: MERGA.

Steinle, V., & Stacey, K. (2004b). Persistence of decimal misconceptions and readiness to move to expertise, *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 225-232). Bergen: PME.

Steinle, V., Stacey, K., & Chambers, D. (2002). *Teaching and Learning about Decimals* [CD-ROM]: Department of Science and Mathematics Education, The University of Melbourne.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

APU	Assessment of Performance Unit (UK)
CSMS	Concepts in Secondary Mathematics and Science (UK)
NAEP	National Assessment of Educational Progress (USA)
TIMSS-R	Third International Mathematics and Science Study (Repeat)
KIM	National Project in Netherlands

CHAPTER 1 INTRODUCTION

Inspecting a mathematics textbook might give the impression that, in order to be competent with decimal numbers, all that students need to do is remember a few rules for placing the decimal point, and otherwise perform operations with decimal numbers as if they were whole numbers. Yet many students do not make this apparently small step with ease. For example, the Third International Mathematics and Science (Repeat) Study conducted in 1999 showed that internationally about only 50% of 13 year old students could select the smallest decimal number from a list of five. Nor does this problem disappear with age. Grossman (1983) found that only 30% of a large sample of students about to commence tertiary study was able to choose the smallest decimal number from a different list of five.

The task of choosing the larger or smaller decimal from a list is important for several reasons. Firstly, this task is an important skill in itself and competency with decimal notation is assumed in many societies. For example, a homebuyer comparing the interest rates at two lending institutions should be able to choose sensibly between 4.8% and 4.35%. Secondly, without any sense of the quantitative value of decimal numbers, students are unable to make sense of the mathematics they encounter in their classrooms. Graeber and Johnson (1991) referred to both of these points,

> Students who lack an understanding of decimals and decimal operations are destined to face serious problems as they compute, perform mental computations and try to solve problems in the middle and high schools and in the “real world”. Such students develop computational procedures that are rote and devoid of any meaning. Confusion reigns as to when to “line up the decimal points”, when to “add” zeroes, when to “count the decimal places” and so on. Further, these students cannot determine whether or not their solutions are correct or even reasonable. (p3-26)

The third reason that this task is of interest is that the pattern of errors on a *decimal comparison test* (consisting of a carefully constructed set of comparison tasks) gives clues to the way in which a person is attempting to make sense of decimal notation. In other words, a test can be constructed that not only allows us to separate those students who are able to correctly complete the task from those who cannot, it further allows us to subdivide students into groups who think in various erroneous ways. Exactly how this is done is discussed in detail in Chapter 3. Hence, the comparison task can be seen as a powerful tool that can be used to uncover students' misconceptions; without this tool, these misconceptions may go undetected and hence unchallenged.

The fact that students harbour misconceptions about what a decimal number actually means has been well established (see for example, Sackur-Grisvard & Leonard, 1985 and Resnick, Nesher, Leonard, Magone, Omanson & Peled, 1989) and furthermore, these problems do not disappear by the time that students leave school. For example, Swan (1990) observes that "the implications of this research are serious…most 15 year old students have no 'feeling' for the relative sizes of decimal numbers at all", (p49) and Hiebert (1987) comments

> For many students, errors made in working with decimal fractions, unlike those made in working with whole numbers, do not correct themselves as students move through school. Low levels of performance on decimal fraction problems are common throughout high school and beyond. (p22)

While some adults might have difficulty with problems involving decimal numbers, the fact that pre-service elementary teachers, in particular, have difficulty is a great concern, (see for example, Thipkong & Davis, 1991, and Putt, 1995). Inspection of the recent medical literature indicates that cases of both overdosing and underdosing of medication are being investigated: for example, *Tenfold medication dose prescribing errors*, (Lesar, 2002) and *The naked decimal point,* (Karch & Karch, 2001). Furthermore, in an interview on the Australian Broadcasting Corporation [Radio National, July 8th, 2001, conducted by Chris Bullock] an investigation into errors by nurses in US hospitals was discussed:

> They found that in a five year period, almost 2,000 people were accidentally killed and almost 10,000 were injured through nurse error. Over 400 people were killed by nurses wrongly programming drug infusion pumps, which regulate the flow of medicine. This kind of calculation error is so prevalent, according to the investigation, that nurses call it 'death by decimal'.

An Australian Research Council funded study "Improving learning outcomes in numeracy: building rich descriptions of children's thinking into a computer-based curriculum delivery system" was conducted under the leadership of chief investigators K. Stacey and E. A. Sonenberg of the University of Melbourne. As part of that study (referred to in this thesis as the ARC Study) a decimal comparison test was created, by adapting and expanding tests used by other researchers; this test was then used to investigate students' misconceptions of decimal notation. On this test, students are asked to circle the larger decimal in each pair of numbers, for example, comparing 0.8 and 0.75. The pattern of answers on the test allows us to allocate the student to a category that indicates how that student is most likely thinking about decimal numbers. While such a link may, at first glance, seem tenuous, earlier research (e.g. Nesher & Peled, 1986, and Resnick, *et al.*, 1989) found that this test format gave reliable evidence of children's thinking, which has been backed up by further written questions as well as interviews, as will be carefully reviewed in Chapter 2.

Several thousand students from 12 Victorian schools were involved in the ARC study, representing a mix of various socio-economic groups and ages (Grades 4 to 10). Repeated testing of these students, over periods of up to four years, provides data on the changes in students' understanding of decimal notation. It is the movement of students, between the various misconceptions, over time that is the prime focus of this thesis.

Earlier longitudinal research on 50 secondary students (Moloney & Stacey, 1997) also using a decimal comparison test, found that many of the students gave the same wrong answers when tested one year later. This thesis will investigate two issues raised in this research. Firstly, the length of time that student's thinking remains unchanged will be explored; this measure of *persistence* could be used as a benchmark against which to compare the effectiveness of intervention strategies in future research. Secondly, the likelihood of the students with various misconceptions becoming expert by the time of their next test will be compared to provide an indication of a *hierarchy amongst misconceptions*. Moloney and Stacey found that, of the students who made errors on their first test, the students most likely to answer as experts one year later were those who did *not* previously hold onto one of the misconceptions. They speculated that holding a fixed misconception might make it more difficult for students to accept new information and move towards expertise. Students making errors who did not clearly demonstrate a fixed misconception may already understand that their ideas are not adequate and thus be searching for a more complete understanding.

We would expect that once a good understanding of the notation is gained, it would be retained for the long term; hence we would expect that once students have answered the test correctly, they should retest in a similar manner. It is possible though, that some students are merely following a recently taught rule and can achieve expert-like results without a full understanding. Such students may then forget the rule and in subsequent tests answer as non-experts. The incidence of this *regression* will be investigated. Sackur-Grisvard and Leonard (1985) used the term regression for students who, during the one testing session, answered "easier" questions correctly (ordering sets of two decimal numbers) but used incorrect rules when faced with more difficult situations, such as ordering sets of three and five numbers. Hence, our use of the term regression is different; students make errors on the same items at different points in time, rather than errors on different items at the same point in time.

Terminology

In this thesis, the term *decimal number* refers to a (base 10) number that is written with a decimal point. It describes the notation in which the number is written, and not the abstract number itself. For example, whilst 0.8, 2.0 and 2.5 are decimal numbers, neither 2 nor 2½ are decimal numbers. In the present usage, there is no restriction on whether the number is greater or less than one. (In Scandinavia, for example, 1.8 is referred to as a decimal *number*, while 0.8 is referred to as a decimal *fraction*, see Brekke, 1996). The author is aware that the term *decimal numbers* is also used more generally to refer to base 10 numbers, including natural numbers, but the alternative of *decimal fractions* was less appealing for two reasons associated with exclusions. Firstly, the term *decimal fraction* has been used to indicate a number with a finite number of decimal places; see for example, National Research Council (2001), p90. As such, the decimal equivalent of one third would not belong to the set of *decimal fractions*. Secondly, the decimal system is able to represent all real numbers (both rational and irrational numbers), so it seems unnecessary to limit the definition to just the set of rational numbers.

Hence, the terms *decimal number* or just *decimal* will be used in this thesis to indicate that a decimal point has been used to write the number, whether or not the number is greater or less than one, and whether it is rational (finite or infinite length) or irrational. Unless specifically mentioned in this thesis, all decimal numbers are of finite length.

The term *ragged* will be used to describe a set of decimals that are not all written to the same number of decimal places. For example, the set {0.8, 0.75} is a set of ragged decimals, while the set {0.80, 0.75} is not. Thus a set of decimals is described as either *equal length* or *ragged*.

In order to assist in the explanations of students' interpretations of decimal numbers, a number such as 64.520 is said to be composed of two portions: the *whole number portion* (64) and the *decimal portion* (520). The *length of a decimal number* is the number of digits explicitly displayed in the decimal portion, so the length of 64.520 is 3.

Decimal numbers such as 0.3 and 1.3 are referred to as *one-digit decimals*, and 0.54, 0.80 and 3.84 are referred to as *two-digit decimals*, etc. A decimal number is said to be *longer than (shorter than)* another if its decimal portion contains more (less) digits. One decimal number is said to be larger (smaller) than another if its numerical value is greater (less) than the numerical value of the other.

The size of two decimal numbers can be compared in several ways. Sometimes, knowledge of properties of the actual numbers involved can be called upon; so that it might be clear that 0.99 is larger than 0.5 because 0.99 is near to 1 and 0.5 is a half. Usually, however, one of two *expert algorithms* is used, with or without understanding. Both algorithms first compare the whole number portions, but operate differently if the whole number portions are the same. Firstly, the *annexe zero algorithm* refers to the procedure of writing zeros on the right end of the shorter decimal until both decimal portions have the same length, and then comparing the decimal portions as whole numbers. So, for example, 0.75 is seen to be less than 0.8 because 75 is less than 80. Secondly, the *left-to-right digit comparison algorithm* refers to the procedure of comparing digits in corresponding columns, moving from left to right until a larger digit is found. So, for example, 0.75 is seen to be less than 0.8 because 7 is less than 8. (Although it is not relevant to this thesis, note that the annexe zero algorithm cannot be applied to infinite decimals and the left-to-right digit comparison algorithm fails for an infinitely repeating string of nines e.g. by falsely predicting 1.4 is greater than 1.39999…..)

Various researchers have used tasks that involve either ordering a set of decimals, or selecting the largest or smallest decimal from a set; in this thesis these will all be referred to as *decimal comparison tasks*. The term *decimal comparison item* will be reserved for pair-wise comparisons of decimal numbers.

In this thesis, the distinction will be made between *behaviour* and the *way of thinking* that causes such behaviour. Previous researchers (Sackur-Grisvard and Leonard, 1985 and Resnick *et al.*, 1989) have referred to "errorful rules" and "misconceptions" but have not clearly made the distinction between behaviour and thinking. Hence, the use of the term *misconception* will be limited to introductory chapters in this thesis and used in a general way; the more precise terms *behaviour* and *way of thinking* will replace it in later chapters.

Additional terms will be defined at the start of the relevant chapter and are contained in the Glossary for future reference.

Outline of thesis

The aim of this thesis is to learn more about children's thinking about decimals. This thesis does not contain, for example, a comparison of the effectiveness of various teaching approaches or resources (such as textbooks, games and concrete materials) to prevent or eliminate such misconceptions. The literature review in Chapter 2 is therefore confined to research about the causes of misconceptions and the tasks that reveal such misconceptions. This historical review highlights the recursive nature of this research. For this reason, this thesis contains a large number of appendices which contain the results of various analyses of the longitudinal data, as well as students' performance on individual test items from this study. It is intended, however, that enough details are provided in the main body of the thesis, so that inspection of the appendices is optional.

Chapter 3 contains details of the two versions of the Decimal Comparison Tests used to collect data in the ARC study. It opens with a discussion of each of the ways of thinking about decimals, with additional links to the research literature. While some of these ways of thinking were discussed in Chapter 2, they are presented in full as the remainder of Chapter 3 explains how codes (intended to represent the children's thinking) are allocated to students' tests.

The methodology in Chapter 4 is presented in three sections: procedures in the ARC study; procedures in this thesis; and *meta-data* (i.e. data about the data). This last section is provided to give an introduction to the data that will help the reader appreciate the complexities that arise in later chapters. Extremely careful analysis has been required to obtain the maximum valid information from the available longitudinal data.

Chapter 5 contains a general analysis of the data in three sections. Firstly, the prevalence of the misconceptions is presented and variations by age and *school group* are investigated. Secondly, students' *test histories* (the full set of tests completed by each student) are introduced and several analyses are presented. Thirdly, students' movements between misconceptions on subsequent tests, referred to as *transitions*, are presented.

Chapter 6 contains an investigation of the issue of *persistence.* How long do students remain in the same category? This is a measure of how "normal" teaching in schools affects the strongly held beliefs of students. Are there some misconceptions that are held onto by students for a longer time?

Chapter 7 contains an investigation of the issue of *hierarchy*. Do student's movements between categories confirm the hierarchy in the misconceptions as suggested in the literature? Amongst students making errors, are those who do *not* hold onto a fixed misconception more likely to become experts, as found by Moloney and Stacey? If it is this group of students who are most likely to become experts, then providing teaching that stimulates students to question their existing beliefs may be extremely important.

Chapter 8 contains an investigation of the issue of *regression.* How often do students appear to regress from expertise? This would give some idea of the proportion of students who learn rules that improve short-term performance in mathematics classes without an adequate supporting understanding of place value, etc, and thus are unable to retain these rules in the long-term. It is our assumption that students would not forget what they "truly understand".

It is expected that throughout this thesis, the analysis of the data will provide evidence that confirms, enhances or contradicts our current understanding of decimal misconceptions. Chapter 9 will focus on each of the misconceptions, co-ordinating results from earlier chapters (e.g. prevalence) with several new analyses to confirm the nature of these misconceptions.

Chapter 10 concludes the thesis, summarising major findings and giving implications for teachers and researchers. Within this chapter, the most useful level of detail (*behaviour* or *ways of thinking*) will be considered. As there are four behaviours, this level of detail is simpler to use than the approximately twelve ways of thinking, yet analysis at this level may actually hide some important details.

Note that all calculations in this thesis have been done to the full accuracy of software, and rounded (typically) to the nearest percent for tables of results.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction and terminology

This chapter contains a first description of decimal misconceptions and traces the evolution of the use of the task of comparing decimals to reveal misconceptions. Additional literature will be discussed in Chapter 3 when a complete list of decimal misconceptions is provided. The seminal research into decimal misconceptions by Swan (1983), Sackur-Grisvard and Leonard (1985), Nesher and Peled (1986) and Resnick *et al*. (1989) formed the basis for much of the research in the following decade. As well as this research, large-scale studies that have used the task of comparing decimals will also be included in this chapter. The details of the tasks used and the *facilities* (percentage correct) for the samples tested will be included.

Two typical comparison items are:

Circle the larger number in the pair (4.8, 4.75)
Circle the larger number in the pair (4.3, 4.65).

One of the most confronting aspects of this research is that students can answer correctly on one item, but incorrectly on another item that *appears* to be addressing the same knowledge or skills. At a superficial glance, both of the items appear to be assessing the same knowledge of decimals and place value. However, a high percentage of students will get one right and the other wrong. Closer examination reveals that students with the first item correct and the second item incorrect have chosen 4.8 and 4.3 and these are the *shorter* decimals in each pair (i.e. they have fewer digits after the decimal point). Likewise, students who chose 4.75 (incorrect) and 4.65 (correct) have chosen the *longer* decimal in each pair (i.e. they have more digits after the decimal point).

The following comment by Swan (1983) reveals the intimate connection between the tasks and the misconceptions: “It is only by asking the right, probing questions that we discover deep misconceptions, and only by knowing which misconceptions are likely do we know which questions are worth asking”, (p65). It will be shown in this chapter how this cycle has been broken to advantage.

One of the difficulties in this chapter is to present information from various countries over the last two decades when researchers have used different tasks to detect misconceptions and different terminology to describe their findings. This is further complicated by assumptions that have been made regarding the *reason* for a student's response. It will be seen that there may be several possible reasons for a student's response, and that researchers have sometimes assumed there was only one. This difficulty will be overcome by adopting the terminology used by Steinle and Stacey (1998a); while the general term *misconception* is useful, these authors have discriminated between the observable *behaviours* and the *ways of thinking* that cause the behaviour. As suggested above, the two main incorrect behaviours exhibited by students, when asked to compare a set of decimal numbers, are:

- Longer-is-larger (L behaviour), choosing the decimal with the *most* digits after the decimal point as the largest, and
- Shorter-is-larger (S behaviour), choosing the decimal with the *fewest* digits after the decimal point as the largest.

While these may seem incredibly naïve groupings, it will be demonstrated below that students do exhibit such behaviours. Note that the L and S behaviours are described throughout this thesis in terms of responses to a task asking for the *largest* (or larger) number; a task asking for the *smallest* (or smaller) number will produce the opposite responses.

While there is considerable research which demonstrates the difficulties that students have with place value tasks with both whole numbers and decimals; with operations with both decimals and fractions; with conversions between fractions and decimals, as well as other tasks which investigate the density of the real numbers (e.g. write a number between two numbers), this chapter does not attempt to systematically review such research. Students' performance on such tasks *will* be discussed, however, when it follows the classification of students into the various misconceptions and therefore provides insights into the nature of the misconceptions.

This chapter also does not attempt to survey the literature on effective methods of teaching about decimal numbers, even though this is integral to other research also associated with the ARC study (e.g. Stacey, Helme, Archer & Condon, 2001) and other publications of the present author (e.g. Steinle, Stacey & Chambers, 2002). The present study is concerned only with the monitoring the progress of students in their normal classes.

In the following discussion of research tasks, lists of decimals are presented in ascending order to reduce the cognitive load for the reader, rather than in the order that they were presented to the students. In a similar vein, a decimal point will be used throughout, even though the original task may have been presented with a decimal comma (the usual notation in many countries).

Section 2.2 contains an overview of the research into the causes of misconceptions. Section 2.3 summarises the early phase of research in which, typically, a single decimal comparison task was used in large-scale studies to assess students' ability to order decimals. This section also includes a discussion of the features of these tasks which make them "easier" / "harder" for students and more or less useful for researchers and teachers.

Research that contributed to the evolution of the Decimal Comparison Test (designed especially to diagnose misconceptions) is then presented in sections 2.4 and 2.5. Section 2.6 reviews the evidence which supports the successful diagnosis of misconceptions. Section 2.7 summarises the literature with respect to the three issues that will be investigated in later chapters; persistence, hierarchy and regression.

2.2 General overview of causes of misconceptions

This overview will provide a brief summary of the literature concerning the causes of misconceptions in general. While references will be made to various decimal misconceptions, these are collected and dealt with more thoroughly in section 3.2.

Before considering the possible causes of misconceptions, several researchers have commented on the *characteristics* of misconceptions. For example, Graeber and Johnson (1991) included the following in a list of characteristics (p 3-15):

- self-evident- one doesn't feel the need to prove them,
- coercive- one is compelled to use them in an initial response,
- widespread among both naïve learners and more academically able students,
- perseverant, very robust, and
- frequently supported by everyday use of language or symbols.

Such a list emphasises the importance of researching misconceptions and how to remove them, if it is not possible to prevent them completely (as claimed by several of the following researchers).

Confrey (1990) reviewed the literature on misconceptions in the three fields of science, mathematics and programming. She noted the varied terms that were in use in these fields; alternative conceptions; student conceptions; pre-conceptions, conceptual primitives; private concepts; alternative frameworks; systematic errors; critical barriers to learning, and naïve theories. She commented (p19) about the early phase of misconceptions research:

> The dominant perspective was that, in learning certain key concepts in the curriculum, students were transforming in an active way what was told to them and those transformations often led to serious misconceptions. Misconceptions were documented to be surprising, pervasive, and resilient. Connections between misconceptions, language, and informal knowledge were proposed.

The suggestion that misconceptions are due to students actively constructing their own ideas is a recurrent theme in the literature. For example, Resnick *et al.* (1989) commented,

> In making these inferences and interpretations, children are very likely to make at least temporary errors. Errorful rules, on this view, are intrinsic to all learning- at least as a temporary phenomenon- because they are a natural result of children's efforts to interpret what they are told and to go beyond the cases actually presented. Several analyses…have shown that these errorful rules are intelligent constructions based on what is more often incomplete than incorrect knowledge. Errorful rules, then, cannot be avoided in instruction. (p26)

Hiebert, Wearne and Tabor (1991) describe how rapid cognitive work by high-achieving students, "may be partially responsible for the myth that understanding always arrives all at once….Low achieving students may build connections more gradually, revealing more of the obstacles, confusions, and partial understandings that accompany students' efforts to become more expert", (p324).

Graeber and Johnson (1991) proposed four ways of grouping misconceptions but acknowledge that these have some overlap (p4-3):

- *overgeneralisation* (if a student takes a concept or procedure that is appropriate for one class and extends it to another class),
- *overspecialisation* (if a students adds some restriction to a concept or procedure that is not characteristic of the entire class),

- *mistranslations* (if a student makes an error translating between words, symbols, tables and graphs) and
- *limited conceptions* (if a student's misconception is traceable to the lack of a concept or procedure).

Various researchers have commented on overgeneralisations as a cause of misconceptions. For example, Streefland (1991) used the term N-distractors for the tendency to associate new (rational number) ideas with the natural numbers. He noted the "powerful suction" of the N-distractors and that "the presence of N-distractor errors in the instructional process for fractions and ratios constitutes a stubborn and unavoidable phenomenon", (p223). Brase (2002) continues in this vein and suggests that whole numbers are *always* the familiar domain or "default setting" (p399) and that fractions and decimals are *always* the novel domain, even for adults. Hence, these researchers believed that the tendency for students to overgeneralise from whole numbers was so strong that they coined a new term to refer to this phenomenon. Several of the decimal misconceptions discussed later this chapter and in Chapter 3, such as *whole number thinking*, are caused by overgeneralising whole number properties.

Students who overgeneralise their knowledge of whole numbers (in particular, that 1-digit whole numbers "live together", followed by 2-digit whole numbers, followed by 3-digit whole numbers, etc.) might think that the 2-digit decimals live together, as do the 1-digit decimals. Whether they then think that the 2-digit decimals are larger or smaller than the 1-digit decimals depends on other aspects of the misconception, as discussed in section 3.2. Bana, Farrell and McIntosh (1997) found that 23% of their sample of Australian students (aged 14) thought there were no decimals between 1.52 and 1.53. The same proportion of students in the Swedish sample gave this response, compared with nearly 1 in 2 students in the USA sample.

Various researchers have noted that students confuse various symbol systems. For example, Swan (1983, p67) noted that some students appeared to confuse the decimal point with the "r" in remainder (9 r 2), the dot in 3.59pm or the comma in the co-ordinate pair (5,2). Markovits and Sowder (1991) interviewed Grade 6 students and found that 12 out of 14 students indicated that they thought 1.4 was the same as ¼. Hiebert (1986, p431) noted this inappropriate "conversion" between decimals and common fractions (students who converted 4/10 to 4.10); similarly Hiebert and Wearne (1986, p209) noted students who converted .09 to 0/9.

Stacey and Steinle (1998) referred to an interview with a pre-service teacher who explained that she chose 0.20 as larger than 0.35 as "I was thinking along a number line and considering decimal numbers to be equivalent to negative numbers. Therefore -20 was larger than -35", (p59).

Hence there is clear evidence that the decimal point has been confused with various other symbol systems; several of the decimal misconceptions (such as *reciprocal thinking* and *negative thinking*) discussed in section 3.2 are due to such confusion. However, these confusions of *symbols* may well relate to deep confusions of *ideas*. Stacey, Helme and Steinle (2001) proposed that the "confusion and interference arrive from the use of the conceptual metaphor of the mirror in three different ways", (p221). Their explanation was based on these three points (p224):

- the natural numbers are the primary elements from which concepts of the other numbers are constructed,
- the metaphor of the mirror is involved in the psychological construction of fractions, negative numbers and place value notation for decimal numbers, although in different ways,
- the observed confusion results from students' merging (confusing or not distinguishing between) the different targets of the same feature of the mirror metaphor under the different analogical mappings.

MacGregor and Stacey (1997) proposed that interference from new teaching might result in misconceptions in algebra. This also applies in the context of decimals. Wearne (1990) observed students who, after instruction on multiplying decimals (including use of a rule for placing the decimal point in the product), then use this new rule inappropriately for addition and subtraction problems. One might predict that teaching which focusses on conceptual understanding, rather than on procedures, is less prone to this phenomenon. Interference of new teaching is discussed further in section 3.2 as the cause of some decimal misconceptions.

MacGregor and Stacey (1997) also showed that another source of algebra misconceptions was misleading teaching materials. These materials may have errors, may be ambiguous or may be limited in the range of contexts, ideas and applications considered. This cause of misconceptions also applies to decimals. For example, Brekke (1996) makes this comment regarding teaching of decimals within the context of money,

> Teachers regularly claim that their pupils manage to solve arithmetic problems involving decimals correctly if money is introduced as a context to such problems. Thus they fail to see that the children do not understand *decimal numbers* in such cases, but rather that such understanding is not needed; it is possible to continue to work as if the numbers are whole, and change one hundred pence to one pound if necessary. It is doubtful whether a continued reference to money will be helpful, when it comes to developing understanding of decimal numbers; on the contrary, this can be a hindrance to the development of a robust decimal concept. (p138)

The decimal misconception referred to as *money thinking* is discussed fully in section 3.2.3. Brousseau (1997) also comments on how the overgeneralisation of whole number properties is supported by teaching which works exclusively with decimals of a fixed length or with decimals always reflecting a money or measurement amount, such as millimetres,

> But in fact, these school decimal numbers [i.e. of a fixed length] are really just whole natural numbers. In every measure there exists an indivisible submultiple, an atom, below which no further distinctions are made. Even if the definition claims that all units of size can be divided by ten, these divisions are never- in elementary teaching- pursued with impunity beyond what is useful or reasonable, even through the convenient fiction of the calculation of a division….Under these conditions, decimal numbers retain a discrete order, that of the natural numbers; many students using this definition will have difficulty in imagining a number between 10.849 and 10.850. (p125)

Sackur-Grisvard and Leonard (1985) commented that students with misconceptions often correctly answered the activities used by teachers and found in textbooks, as they involve equal length decimals,

> It appears that, over time, teachers and exercise books adapt themselves to students by avoiding problems that are too difficult. By including a majority of problems that can be solved correctly by children, high rates of success in school can be assured. (p171)

Hence, a *staged* introduction to decimals, in which only equal length decimals are considered, is unlikely to provide students with suitable experiences for them to appreciate *number density*. Whereas there is no whole number between 3 and 4, there is always a decimal number between any two unequal decimals. Students' incorrect beliefs about decimals are often exposed when they are asked to insert a number between two *apparently consecutive* numbers, for example, between 0.3 and 0.4, or between 0.52 and 0.53.

Examination of students' completed Decimal Comparison Tests (described fully in Chapter 3) provides evidence of misleading or limited teaching. While the *annexe zero algorithm* provides correct answers, some students wrote over 40 zeros on their test paper. Swan (1990) comments on the use of this procedure, "This rule (which to many will seem arbitrary and meaningless) provides correct answers, but will not help to remove any of the misconceptions. It may even perpetuate them", (p49). Hence, teachers who provide this procedure to their students, instead of emphasising place value, are providing limited teaching which is unlikely to assist students to improve their conceptual understanding of decimal notation and may mask or cause, rather than overcome, decimal misconceptions.

Brown (1981) reported on the Concepts in Secondary Mathematics and Science (CSMS) project in the United Kingdom in the late 1970s, which included large-scale testing as well as interviews. In these interviews students were asked to multiply 5.13 by ten. A first attempt by some students was to use the *add a nought rule* and obtain 5.130. While this is an overgeneralisation of a rule for whole numbers, the fact that some students were not content with this answer and rejected it, leads to another observation. Students unhappy with 5.130 modified their rule to obtain the answer 50.130. This is an example of students inventing a rule to deal with a situation where their current rule fails to provide them with a satisfactory answer. Such "repairs" to procedures are now discussed.

The now famous work on *buggy algorithms* by Brown and VanLehn (1982) considered the systematic errors (bugs) made by students as they performed multi-digit subtraction. They explain how such bugs may be the result of a repair to impoverished procedures as follows,

> The theory is motivated by the belief that when a student has unsuccessfully applied a procedure to a given problem, he or she will attempt a repair. Suppose he or she is missing a fragment (subprocedure) of some correct procedural skill, whether because he or she never learned the subprocedure or maybe forgot it. Because the missing fragment must have had a purpose, attempting to follow the impoverished procedure rigorously will often lead to an impasse….he or she will often be inventive, invoking problem-solving skills in an attempt to repair the impasse and continuing to execute the procedure, albeit in a potentially erroneous way. We believe that many bugs can best be

> explained as patches derived from repairing a procedure that has encountered an impasse whilst solving a particular problem. (p122)

Brown and VanLehn's discussion of the phenomenon of "bug migration" is relevant to this thesis,

> Students have been observed who had a bug one day and a different bug a few days later... Suppose that on the first occurrence of an impasse on a test, the student created a repair and used it throughout the test…..Suppose further that when he or she takes the second test, a few days later, he or she has forgotten the repair made previously. Hence, when an impasse is encountered, he or she may use a different repair and hence exhibit a different bug. This is the theory's explanation of bug migration. However, it also makes a prediction. It predicts that bugs that migrate into each other will be related in that they are different repairs to the same impasse. (p135)

Due to the longitudinal nature of the data being analysed in this thesis, it may be possible to determine which decimal misconceptions are related if they are repairs to the same impasse. Some students complete the Decimal Comparison Task by trying to follow taught rules, and evidence for bug migration may be identified amongst these students.

2.3 The early phase of research

Researchers have been using the decimal comparison task for many years, for example, Brueckner (1928); Brueckner and Bond (1955); and Af Ekenstam (1977). The next three sections (2.3, 2.4 and 2.5) will describe the research by which decimal misconceptions came to be known, especially research that used decimal comparison tasks. In this section, details of large-scale studies that included this task will be provided to demonstrate that firstly, comparing decimals is not as easy as some might expect, and secondly, that some tasks are *better* than others. This leads to a discussion of the features that make a task easier or harder for students and more or less useful for researchers and teachers.

2.3.1 Large-scale studies using decimal comparison tasks

Brown (1981) reported on the decimal comparison tasks used in the Concepts in Secondary Mathematics and Science project in the United Kingdom, which was followed by the Assessment of Performance Unit (APU) reported by Foxman, Ruddock, Joffe, Mason, Mitchell, and Sexton (1985). Carpenter, Corbitt, Kepner, Lindquist, and Reys (1981) reported on the results of the second National Assessment of Educational

Progress (NAEP2) in the USA, and Kouba, Brown, Carpenter, Lindquist, Silver and Swafford (1989) provided the results of NAEP4.

Grossman (1983) reported on tasks from two versions of an entrance examination to a university in New York. Over 7000 students, in total, were involved. Not only did these decimal comparison tasks have low facilities (30%), *more* students (40%) chose the longest decimal in the given set, (corresponding to S behaviour - see section 2.1), than chose correctly! A similar result was found by Putt (1995) who tested approximately 700 university students involved in pre-service teacher education. While about one half of the pre-service teachers answered correctly, 36% listed 0.6 as the largest number (corresponding to S behaviour).

Brekke (1996) reported on a large-scale study (KIM Project) of over 1500 Scandinavian students in Grades 4, 6 and 8 (ages 11, 13 and 15), and found marked progress across the age groups, but also generally low facilities (i.e. percentage correct). Fuglestad (1998) also tested Scandinavian students; over 600 students from Grades 5 to 7, (ages 10-14), and found a similar pattern of results. An intervention program, based on spreadsheets, was used to generate conflict for over 200 students with misconceptions.

Hiebert and Wearne (1986) reported on the results of a study of about 700 students from Grades 4 to 9 over 2 years involving both written tasks and interviews and included a comparison task (choose the largest decimal from a list of 4).

The Third International Mathematics and Science Repeat Study (TIMSS-R) conducted in 1999 showed that, internationally, about only 50% of 13 year old students could select the smallest decimal number from a list of five (item B10).

The details of the decimal comparison tasks and the facilities for the sample involved in these studies are provided in Table 2.1. The facilities are given by likely average age of sample, although precise data is not available in all cases. The purpose of this table is not to compare the performance of students in the various countries, but rather to make the point that the task of ordering decimals is not as simple as many expect. Furthermore, it shows that slight changes to a task (for example, changing the instruction from *largest* on task 4 of the APU study, to *smallest* to create task 5) made an incredible difference to the number of students choosing correctly (dropping from

82% to 47%). Note that if these tasks involved only a pair of decimals, then such changes to the instruction would not have produced this effect.

Table 2.1: Facilities for various tasks in large-scale studies in 1980s and 1990s

Researcher and details of task		Approximate Age					
(order a list, select smallest, select largest)		11	12	13	14	15	15^+
Brown (1981) CSMS							
1) Larger	0.75 0.8		57	65	69	75	
2) Larger	4.06 4.5		66	72	83	80	
Carpenter et al. (1981) NAEP2							
1) Larger	0.23 1.9			81			
2) Larger	1.15 1.36			79			
3) Largest	.036 .19 .195 .2			46			
Grossman (1983)							
1) Smallest	0.07 0.075 0.08 0.3 1.003						29
2) Smallest	0.004 0.03001 0.05 0.1 1.0003						31
Foxman et al. (1985) APU							
1) Order	0.07 0.1 0.23	23					
2) Order	0.1 0.3 0.6 0.7	75					
3) Smallest	0.125 0.25 0.375 0.5	17					
4) Largest	0.075 0.089 0.09 0.1					82	
5) Smallest	0.075 0.089 0.09 0.1					47	
6) Largest	0.125 0.25 0.375 0.5 0.625					61	
7) Smallest	0.125 0.25 0.375 0.5 0.625					37	
8) Smallest	0.125 0.25 0.3753 0.5 0.625					43	
Hiebert & Wearne (1986)							
Largest	.09 .1814 .3 .385	0				43	
Kouba et al. (1989) NAEP4							
Largest	0.058 0.36 0.375 0.4		50*				
Putt (1995)							
Order	0.060 0.0666 0.6 0.606 0.66						51
Brekke (1996)							
1) Largest	3.521 3.6 3.75	20		64		88	
2) Largest	0.649 0.7 0.87	22		62		83	
3) Smallest	0.125 0.25 0.3753 0.5 0.625	16		55		79	
Fuglestad (1998)							
Order	0.25 0.375 0.5 0.62	20	35	62			
TIMSS-R (1999)							
Smallest	0.125 0.25 0.375 0.5 0.625			46			

* *approximate*

In fact, it was as a result of the comprehensive analysis of the responses to the APU tasks and the effect of slight changes to either the instructions or to one of the distractors (as discussed in the next section) that Foxman *et al.* confirmed the existence of two large groups of students with misconceptions (referred to in this thesis as L and S behaviours; see section 2.1). They note that this S behaviour (their term *largest is smallest*) was unexpected,

> Despite the large proportions of pupils giving this type of response very few teachers, advisors, and other educationalists are aware of its existence – the monitoring team were among those unaware of the 'largest is smallest' response at the beginning of the series of surveys. (p851)

2.3.2 Characteristics of a *good* decimal comparison task

There are various types of tasks in Table 2.1 (order a list, select smaller/smallest and select larger/largest) and also variations in the number of decimals provided. What are the features of a *good* comparison task?

First consider the three tasks reported by Carpenter *et al.*; tasks 1 and 2 have a facility of about 80%, while the facility of task 3 is less than 50%. What features of the tasks make them easier or harder for students? In this case, task 3 contains a set of four decimals, while the other tasks have only two decimals. Since there are more cognitive steps required to compare four than two decimals, this is likely to be one reason for the lower facility; there are simply more steps where an error can be made. This result was demonstrated in the data of Sackur-Grisvard and Leonard (see section 2.7.3). Furthermore, task 3 contains longer decimals than tasks 1 and 2 (i.e. 3-digit instead of 1-digit and 2-digit). Recently, Steinle and Stacey (2003a), reanalysing the data used in this thesis, have confirmed this to be a background factor affecting students' performance (see section 2.5).

Another factor that explains some variation in facilities, although it does not apply in the case of the Carpenter *et al.* tasks, is whether the question asks for the shorter/shortest or longer/longest. Information processing theory has established over a wide range of situations that *marked terms* such as "smaller" are processed in the brain with reference to the corresponding *unmarked term*, in this case "larger" (Clark & Clark, 1977). To decide which of two decimals is smaller, this general principle indicates that the item is understood to be the one that is *not larger*, and thus additional processing is involved, with consequently more possibility for error.

The major factor, however, that explains the decrease in facility is the exact nature of the numbers in the tasks and this is now discussed.

Where there are two decimal comparison tasks with different facilities (for a particular sample) then it is likely that students with misconceptions, who have chosen the correct answer for the wrong reason, inflate the higher facility. In other words, the task with lower facility is not so much a *harder* task as a *better* task, as students with misconceptions are separated from students able to compare correctly in all situations. Hence, the major reason that the APU task 4 has a higher facility than task 5 (82% compared with 47%), is that students with misconceptions are choosing the correct answer in task 4.

That this must be the case can be seen in the data reported by Vance (1986). Over 130 Grade 6 and Grade 7 Canadian students completed various tasks; the facilities for the two samples of students on four particular tasks are provided in Table 2.2. The facilities range from 27% to 85% for the Grade 6 students, and 44% to 75% for the Grade 7 students. This range of facilities can be explained by noting that the tasks with the lower facilities have answers with two decimal places (0.47 is the smallest number in set 1, and 0.32 is the largest number in set 2).

Table 2.2: Facility on four tasks by Vance (1986)

	Grade 6 (n=62)	Grade 7 (n=69)
Set 1: 0.47 0.5 0.613		
Largest	85	75
Smallest	27	48
Set 2: 0.3 0.302 0.32		
Largest	32	44
Smallest	77	69

Foxman *et al.* referred to these tasks with lower facility as *true* tests of the ability to compare decimals, as they provided appropriate distractors to separate the two groups of students with misconceptions that they had identified.

The facility for the TIMSS-R task listed in Table 2.1 (46%) is the international average; Australian students were more successful on this task (58%). Table 2.3 contains the percentage of students in both the international sample and the Australian sample (data held at Australian Council of Educational Research, Melbourne) who chose each of the five options. As this same task was used in the APU study reported by Foxman *et al.*, this distribution of students is also included in this table.

More students in each sample chose the correct answer of 0.125 than any other option. There are two distractors that were chosen by substantial numbers of students in each sample: 0.5 would be chosen by students who were exhibiting L behaviour, and 0.625 would be chosen by some of the students who were exhibiting S behaviour. Unfortunately, the remainder of the students who were exhibiting S behaviour are most likely to choose 0.125, which is the correct answer.

In other words, this task splits the students exhibiting S behaviour into two groups: those who think that 0.625 < 0.125 (referred to in this thesis as *reciprocal thinking* and *negative thinking*) and those who think that 0.125 < 0.625 (referred to in this thesis as *denominator focussed thinking* and *place value number line thinking*). These ways of thinking are explained fully in Chapter 3.

Table 2.3: Comparison of the distribution of students from TIMSS-R and APU studies on one task: *Which of these is the smallest number?*

Option	TIMMS-R age 13		Foxman *et al.*	Comment
	International	Australian	APU, age 15	
0.125	**46**	**58**	**37**	Correct, some S behaviour[a]
0.25	4	4	3	
0.375	2	1	2	
0.5	24	15	22	L behaviour
0.625	24	22	34	some S behaviour[b]

a: denominator focussed thinking and place value number line thinking
b: reciprocal thinking and negative thinking

An “apparently” similar task is now discussed which illustrates the importance of considering student misconceptions in the creation of comparison tasks. A “slight” change to the APU task 7 (writing the digit 3 at the end of one of the distractors) created task 8 (see Table 2.1). While the facilities of both tasks are similar (about 40%), the effect on the students choosing various distractors was enormous.

Table 2.4 provides the distribution of students choosing the various options reported by Foxman *et al.*, as well as those reported by Brekke, for three different age groups. Table 2.4 indicates that the distractor 0.625 was chosen by only 1-4% of the students, compared with about 20-30% of the students in Table 2.3. Similarly, Table 2.4 indicates that the distractor 0.3753 was chosen by 36% of the 15 year olds in the APU study, compared with only 1-2% of students choosing 0.375 in Table 2.3.

Table 2.4: Comparison of the distribution of students from KIM and APU studies on one task: *Which of these is the smallest number?*

Option	Brekke, KIM project			Foxman *et al.*	Comment
	age 11	age 13	age 15	APU, age 15	
0.125	16	55	79	43	Correct
0.25	11[a]	5[a]	3[a]	2	
0.3753	8	13	10	36	S behaviour
0.5	64	26	7	13	L behaviour
0.625	1	1	1	4	

a: Approximate figures due to typographical error in original paper

Swan (1983) used a single cleverly designed task to diagnose decimal misconceptions in 12 year old students in an English school: *Ring the BIGGER number* from the list 0.236, 0.4, 0.62. A student exhibiting L behaviour would select 0.236, a student exhibiting S behaviour would select 0.4 and a student following a correct algorithm would choose 0.62. The correct answer to this task is *neither* the longest *nor* the shortest decimal provided. Table 2.5 provides the distribution of students choosing each option in Swan’s sample, as well as the results for Greer (1987) and Fuglestad (1998) who used the same task in their research. Brekke (1996) and Hiebert and Wearne (1986) used similar tasks and these are also included in this table.

The tasks in Table 2.5 demonstrate that it is possible to separate students exhibiting L or S behaviour from those who do not, with the use of a *single* cleverly designed task. This requires the correct answer to be neither the longest nor the shortest decimal in the set. As students rarely encounter decimals of length four or five, such a task would typically have decimals of length one, two and three. It is considerably limiting, however, for the correct answer on every such task to always be a two-digit decimal! As an alternative, Swan (1983) indicated how a sequence of tasks could be used to separate students with three misconceptions; see Table 2.6. Note the predictions for the misconception *ignore decimal point*; such a student would choose 6.78 as larger than 345, as 678 is larger than 345.

Note that one of the variations of L behaviour discussed in section 3.2.1 (*string length thinking*) would choose 4.008, rather than 4.09, in task 4.

Table 2.5: Comparison of the distribution of students from five studies on one task: *Ring the bigger number*

Researcher	Country	Age Grade	Sample size	*L*	*S*	*Correct*
Task 1				**0.236**	**0.4**	**0.62**
Swan (1983)	England	age 12	(n=98)	50	28	17
Greer (1987)	Ireland	age 12/13	(n=65)	12	43	45
Fuglestad (1998)[a]	Norway	Grade 5	(n=220)	60	7	31
Fuglestad (1998)[a]	Norway	Grade 6	(n=216)	42	6	50
Fuglestad (1998)[a]	Norway	Grade 7	(n=200)	19	4	77
Task 2				**.1814**	**.3**	**.385**
Hiebert *et al.*(1986)[b]	USA	Grade 5	NA	89	6	0
Hiebert *et al.*(1986)[b]	USA	Grade 9	NA	31	25	43
Task 3				**3.521**	**3.6**	**3.75**
Brekke (1996)	Norway	age 11	national	74	5	20
Brekke (1996)	Norway	age 13	national	30	6	64
Brekke (1996)	Norway	age 15	national	6	5	88
Task 4				**0.649**	**0.7**	**0.87**
Brekke (1996)	Norway	age 11	national	66	8	22
Brekke (1996)	Norway	age 13	national	26	10	62
Brekke (1996)	Norway	age 15	national	7	9	83

a: between 1% and 3% chose a fourth option "I don't know"
b: between 1% and 5% chose a fourth option .09
NA: not available, over 700 students in a longitudinal study from Grades 4 to 9

Table 2.6: Patterns of responses used to detect misconceptions, Swan (1983), p73

Tasks (choose largest)	Ignore decimal point	L behaviour[a]	S behaviour[b]	Correct
1) 6.78, 45.6, 345	6.78	345	345	345
2) 3.521, 3.6, 3.75	3.521	3.521	3.6	3.75
3) 15.327, 15.4, 15.56	15.327	15.327	15.4	15.56
4) 4.008, 4.09, 4.7	4.008	4.09[c]	4.7	4.7

a: Swan's term separator
b: Swan's term longer-is-smaller
c: Not string length thinking

Hence, we see that the pattern of responses on a sequence of carefully constructed decimal comparison tasks can be used to separate students with various misconceptions.

One further note on the features of a *good* comparison task. The research by Resnick *et al.* (1989) is discussed fully in the next section, but is mentioned briefly here. Interviews with students revealed that comparison items that involved one decimal and one common fraction were likely to result in inconsistent responses by students. These items have since been discarded from subsequent versions of decimal comparison tests (e.g. Moloney & Stacey, 1997) so that lack of knowledge of common fractions is not a confounding factor in the classification.

Hence, one version of a *good* comparison task is to ask students for the largest number from a set of decimals, where the correct answer is neither the longest nor the shortest number. A more flexible version is to provide a sequence of decimal comparison items (consisting only of pairs of decimals), with the instruction to select the larger number in each pair. These items must separate the two groups of students exhibiting misconceptions (L behaviour and S behaviour) from students who use an expert algorithm.

In the following sections, a series of studies that systematically explored the various misconceptions and attempted to probe their origins will be discussed. These studies are characterised by the use of a sequence of decimal comparison tasks that will be referred to as *decimal comparison tests*. For each version of a decimal comparison test, a *classification system* is required which indicates the predicted responses for students with misconceptions on the various tasks within the test, and describes the connection between response patterns and diagnosis.

2.4 Four rules: WNR, ZR, FR, ER

In this section, a series of publications based on the same four misconceptions (known as rules in the papers) will be discussed; Sackur-Grisvard and Leonard (1985), Nesher and Peled (1986), Resnick, Nesher, Leonard, Magone, Omanson and Peled (1989), Moloney (1994), Baturo and Cooper (1995), Moloney and Stacey (1996, 1997), and Stacey and Steinle (1998). While Sackur-Grisvard and Leonard provided comparison tasks with three decimal numbers, the remaining researchers have used pairs of decimals and the term *decimal comparison item,* or just *item*, will be used to indicate such a pair-wise comparison task.

With regard to the grouping of students, the terminology used in this section, and explained below, will be that of Resnick *et al*. (1989); that is, Whole Number Rule, (WNR), Fraction Rule (FR), Zero Rule (ZR) and Expert Rule (ER). Students not allocated to one of these four rules will be labelled as Unclassified (UN). (Note that both Whole Number Rule and Zero Rule correspond to L behaviour and Fraction Rule is one of the explanations for S behaviour).

With regard to the grouping of tasks, the terminology used in this section will be that of Stacey and Steinle (1998); that is, items are grouped into Type A (e.g. 4.63 / 4.8), Type B (e.g. 4.08 / 4.7) and Type C (e.g. 4.45 / 4.4502). Note that these item types are defined according to the predicted responses to students with various misconceptions, as in Table 2.7. The *classification system* is then a table matching item types and predicted responses from students using different rules; see Table 2.7. Students using any given rule should answer consistently (i.e. correctly or incorrectly) to all items of the same type. In practice, an allowance is made for a small number of "careless responses" (i.e. deviations from the predictions).

Table 2.7: Classification system adapted from Resnick *et al*. (1989)

Item Types	Sample item (larger listed last)	Rule usage				
		ER	WNR	FR	ZR	UN
A	4.63 / 4.8	√	X	√	X	
B	4.08 / 4.7	√	X	√	√	else
C	4.45 / 4.4502	√	√	X	√	

While Expert Rule was allocated to students who made almost no errors on the items within the particular decimal comparison test being used, (see Table 2.7) the other rules are now explained. Whole Number Rule users would incorrectly choose 4.63 as larger than 4.8 (as 63 > 8); incorrectly choose 4.08 as larger than 4.7 (as 8 > 7) and correctly choose 4.4502 as larger than 4.45 (as 4502 > 45).

The Whole Number Rule was explained as students interpreting the decimal portion of the number as a whole number. Various suggestions were made for the reason behind the Fraction Rule. For example, Resnick *et al.* suggested that as two-digit decimals represent hundredths and one-digit decimals represent tenths, then a student who focussed on the size of the parts "might well infer that longer decimals, because they refer to smaller parts, must have lower values", (p13).

Note that these researchers did not look for Swan's *ignore decimal point* misconception, as it was predicted to be fragile and short-lived. Hence, all items within Types A, B and C have whole number parts equal; a student who ignored the decimal point would then be allocated to the Whole Number Rule.

Nesher and Peled explain the Zero Rule as students who normally use Whole Number Rule and then, in the case of zero in the tenths column (e.g. 4.08 and 4.7), "change their rule and say …that 0 in the tenths makes the number always smaller", (p70). These students had added an extra piece of information to their otherwise unchanged misconception.

Baturo and Cooper (1995) reported on 130 Grade 5 students who completed a written test (containing nine comparison items with a maximum of two decimal places) and then follow-up interviews. They considered the four rules above, (defining the Expert Rule as the explicit use of the left-to-right digit comparison algorithm) and then five additional strategies were identified:

renaming	renaming tenths as hundredths
benchmarking	based on estimates, eg less than one half
zero-ignored	4.08 > 4.7 as 8 > 7
expert-backward	compare like places right to left
fraction-inverted	find the missing numbers of tenths and hundredths to make a whole, but then ignoring the size of the parts and comparing these two whole numbers

Note that the *renaming* and *benchmarking* strategies would be expected to produce correct answers. As the annexe zero algorithm is not mentioned explicitly, it is possible that students who used this algorithm were assumed to be *renaming* tenths as hundredths, although their thinking could have been much less sophisticated. The *zero-ignored* strategy is likely to be the same as *numerator focussed thinking*, see section 3.2.1.

Moloney and Stacey (1997) used a decimal comparison test consisting of 15 items (see Appendix 1) to diagnose students' misconceptions. In their first study, they tested 50 students on two occasions; firstly when the students were in Grades 7 and 9 and again one year later when these students were in Grades 8 and 10. Their second study involved testing over 350 students from Grades 4 to 10 to determine the prevalence of the various rules. Furthermore, additional tasks were used to probe the knowledge of the students assigned to the various rules. The results of both of these studies are discussed later this chapter.

2.4.1 Prevalence of four rules

In their first study, Sackur-Grisvard and Leonard provided students with sets of three decimals to order; see Table 2.8. For task 1, students using Zero Rule would be expected to give the same reply as students using Whole Number Rule. For task 2, students using Fraction Rule would be expected to give the correct answer, and hence inflate the facility on this task. Task 3, however, separates all the incorrect rules from Expert Rule. For this reason, the distribution of students by grade on task 3 only will be provided in Table 2.9 (rather than the original estimates provided by the researchers, which were based on averaging the results from various tasks, a process which is not justified). Table 2.9 contains a summary of the prevalence of the various rules as found by this group of researchers.

Table 2.8: Predicted ordering of decimals according to rule use, Sackur-Grisvard & Leonard (1985)

Task	ER	WNR	FR	ZR
1	3.53 3.682 3.7	3.7 3.53 3.682	3.682 3.53 3.7	Same as WNR
2	7.087 7.65 7.8	7.8 7.65 7.087	Same as ER	7.087 7.8 7.65
3	6.07 6.296 6.4	6.4 6.07 6.296	6.296 6.07 6.4	6.07 6.4 6.296

There are several reasons that a detailed discussion of the variations in prevalence of the various rules between the different studies will not be attempted. For example, there are variations in the number of items used in the various decimal comparison tests, as well as their nature (for example, Nesher and Peled included items with a common fraction and a decimal) and convenience sampling has generally been used. Furthermore, in some studies, the evidence from the decimal comparison test was supplemented by interviews to assist in the allocation of students to rules. Even when there was consistency in the test used, for example across the grades in the Moloney and Stacey study, there can be considerable variations in the prevalence of the rules, which is likely to be explained by sampling.

Table 2.9: Prevalence of four rules: various researchers and grades

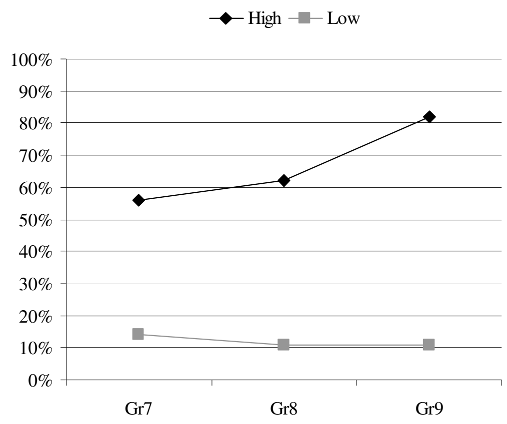
Details of study		Rule Usage				
		ER	WNR	FR	ZR	UN
Sackur-Grisvard & Leonard (1985)						
France Grade 4	(n=52)	42	38	2	8	10
France Grade 5	(n=49)	51	27	4	14	4
France Grade 6	(n=57)	58	16	5	11	11
France Grade 7	(n=69)	74	9	1	10	6
Nesher & Peled (1986)						
Israel Grade 6	(n=21)	19	19	33	14	14
Resnick et al. (1989)						
USA Grade 5	(n=17)	18	35	18	0	29
Israel Grade 6	(n=21)	19	19	33	14	14
France Grade 4	(n=37)	30	41	8	11	11
France Grade 5	(n=38)	53	18	3	24	3
Baturo & Cooper (1995)						
Australia Grade 5	(n=130)	34	25	6	8	17*
Moloney & Stacey (1997)						
Australia Grade 4	(n=60)	22	3	60	0	15
Australia Grade 5	(n=52)	8	42	33	0	17
Australia Grade 6	(n=59)	41	19	25	3	12
Australia Grade 7	(n=58)	31	7	43	14	5
Australia Grade 8	(n=56)	64	4	28	0	4
Australia Grade 9	(n=49)	69	2	27	2	0
Australia Grade 10	(n=45)	73	0	20	4	2

** approximate*

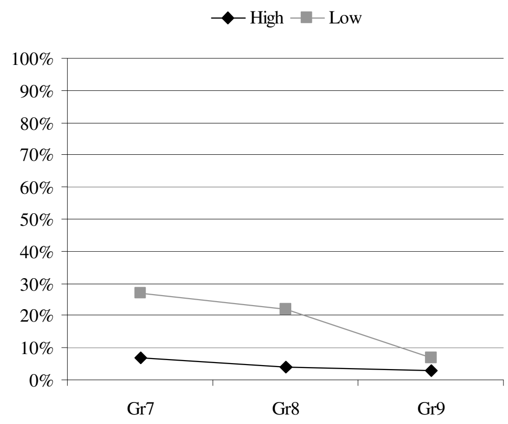
Some general comments, however, can be made. Not surprisingly the prevalence of the Expert Rule typically increases with age and is accompanied by a decrease in the prevalence of the Whole Number Rule. Typically, there are no such trends with age/grade with either Fraction Rule or Zero Rule, although Fraction Rule is clearly more prevalent than Zero Rule except in France, which Resnick *et al.* attribute to a curriculum sequence which does not teach decimals from fractions as is done in USA and Israel.

Nesher and Peled (1986) conducted two studies; the first involved interviewing Grade 6 students (see Table 2.9 for prevalence of rules), and in the second study, 240 students in Grades 7 to 9 completed a 30-item decimal comparison test. The prevalence of the rules for this larger sample was provided for two ability levels within each grade; presumably ability was assessed on other tests. Figure 2.1 has been created from the prevalence of the various rules provided by Nesher and Peled in Table V (p77).

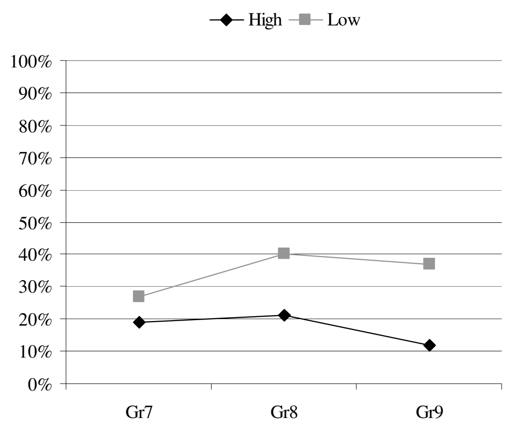
As expected, the prevalence of Expert Rule increases with grade for the high ability students; see Figure 2.1a. What is rather surprising is the lack of increase in the prevalence of Expert Rule in the low ability students; only 1 in 10 of these Grade 9 students was allocated to Expert Rule. What happens to these older, low ability students? Figure 2.1b shows that they do not use Whole Number Rule or Zero Rule, while Figure 2.1c shows that the prevalence of Fraction Rule among the low ability students increases from Grade 7 to 8 (up to 40%) and then remains at just below 40% in Grade 9. Figure 2.1d shows that over 40% of these low ability students in Grade 9 made errors, but were unable to be allocated to one of the rules. Hence, while it is pleasing that fewer of these older students are being allocated to either Whole Number Rule or Zero Rule, they are not all becoming experts; rather they are using Fraction Rule or making errors that do not fit predictions.



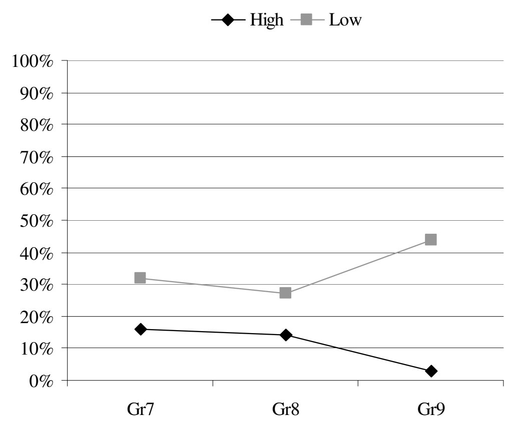
a) Prevalence of ER students



b) Prevalence of WNR/ZR students



c) Prevalence of FR students



d) Prevalence of UN students

Figure 2.1: Prevalence of various rules by grade for high and low ability students, Nesher and Peled (1986), p77

2.4.2 Limitations of the four rules

Stacey and Steinle (1998) analysed over 1800 student's responses to a 25-item decimal comparison test (referred to in this thesis as DCT1 and contained in Appendix 1) created from Moloney and Stacey's 15 items (also contained in Appendix 1) with an additional 10 items to probe further. Table 2.10 provides the item analysis from this publication (Table 5, p57) to aid the following discussion.

If no deviations were allowed from the predictions of correct and incorrect on each of these items in order to classify students, then the facilities for the core items (Types A, B and C) under the columns Expert Rule, Whole Number Rule, Zero Rule, and Fraction Rule would be either 0 or 100. A deviation of at most one per item type was allowed, however, to ensure that this hurdle was not set too high. In other words, while Whole Number Rule users are expected to choose the *incorrect* answer on each of the five Type A items, they were allowed to answer just one of these five items correctly and still be allocated to Whole Number Rule. Hence, while it is expected that every Whole Number Rule user would answer item A1 (4.63 / 4.8) incorrectly, Table 2.10 shows that 5% of the Whole Number Rule users answered correctly. Shading is used in Table 2.10 to indicate the entries that are 10% or more from the extreme values of 0% and 100%.

Table 2.10: Facility of DCT2 items by rule use, Stacey and Steinle (1998), p57

		Item	ER (n=563)	WNR (n=457)	ZR (n=66)	FR (n=295)	UN (n=472)
Core Items: (used in classification)	A1	4.63 / 4.8	98	5	24	95	63
	A2	0.36 / 0.5	99	2	6	95	52
	A3	0.75 / 0.8	98	1	2	95	53
	A4	0.216 / 0.37	98	2	9	97	54
	A5	0.100 / 0.25	93	4	15	94	48
	B1	4.08 / 4.7	98	13	79	94	69
	B2	2.0687986 / 2.621	97	2	88	99	65
	B3	3.073 / 3.72	97	4	98	96	71
	B4	8.052573 / 8.514	98	3	86	97	67
	C1	4.45 / 4.4502	93	97	97	4	56
	C2	17.35 / 17.353	98	99	94	1	59
	C3	8.245 / 8.24563	98	99	88	1	58
	C4	0.4 / 0.457	99	99	98	5	66
	C5	5.62 / 5.736	100	99	95	18	70
Supplementary Items: (not used in classification)	D1	0.5 / 0.75	97	98	95	23	75
	D2	0.25 / 0.5	91	3	15	88	53
	D3	0.3333333 / 0.99	98	15	68	92	74
	E1	0.3 / 0.4	99	98	95	59	87
	E2	1.84 / 1.85	99	95	95	54	84
	F1	0.006 / 0.53	98	67	98	93	84
	F2	0.021 / 0.21	99	26	91	92	77
	F3	6.01 / 6.1	99	23	88	94	75
	G1	1.053 / 1.06	96	3	27	93	57
	G2	0.038 / 0.04	91	4	17	89	48
	G3	2.0053 / 2.06	98	6	76	97	70

Darker *shaded cells indicate inconsistent responses in core items (used to classify)*
Lighter *shaded cells indicate inconsistent responses in supplementary items*

This allowance of one deviation (per item type) from the predicted responses was initially adopted as being realistic if students made an occasional random choice, yet quite stringent (earlier studies used "most" as the criteria). The data that was generated by this allowance of one deviation (per item type) provided evidence of the limitations of the current system of classifying students into misconceptions as explained below.

The major finding of the item analysis by Stacey and Steinle was that students within Fraction Rule were not behaving as one group. The facilities of approximately 50% on two supplementary items, E1 (0.3 / 0.4) and E2 (1.84 / 1.85), suggested that there were at least two distinct subgroups within Fraction Rule. Furthermore, the unusually high facility of 18% for the Fraction Rule students on item C5 (5.62 / 5.736) indicated that this item was not eliciting the same response as the remaining Type C items; in other words, the five Type C items were not *homogeneous*.

This analysis indicated that both the current grouping of items and the current grouping of students were inadequate. For this reason a new test (DCT2), which provided most of the data for this thesis, was constructed. Stricter definitions of the item types were made, splitting them to make 6 types, described in Appendix 1. In addition, supported by interview evidence, several new misconceptions were reported, relating to Types 4, 5 and 6 as described in the next chapter.

2.5 Four behaviours A, L, S, U and twelve ways of thinking

A series of papers based on the 30-item test used in the present study (referred to as DCT2) have been published; the results of two of these papers will be discussed below (Steinle & Stacey, 1998a and 2003a). Four papers that involve preliminary analyses of the longitudinal data being analysed in this thesis (Stacey & Steinle, 1999a and 1999b, and Steinle & Stacey, 2002 and 2003b) will be considered in later chapters.

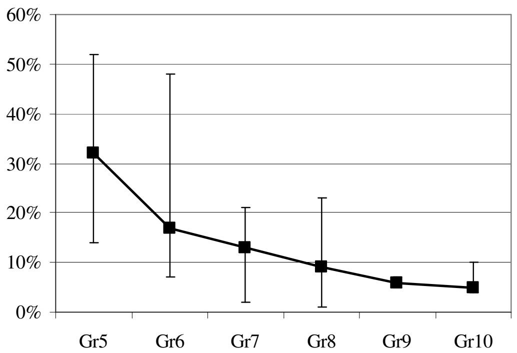
As mentioned in the introduction to this chapter, the "best" classification of student misconceptions (at this point in time) uses two levels. The four coarse-grained *behaviours* (A, L, S and U) are allocated on the basis of the responses to ten items on the test (referred to as Types 1 and 2). As before, L behaviour indicates *longer-is-larger* and S behaviour indicates *shorter-is-larger*. A test which has almost no errors on these ten items is allocated the code A, and U indicates a test which is not allocated A, L or S. Precise definitions are given in section 3.2.

The second level of classification is a subdivision of A, L, S and U into twelve fine-grained *ways of thinking* (e.g. A1, A2, A3) by considering the pattern of responses on further test items (called Types 3 to 6 – see section 3.3 for details of the allocation of codes to students' test papers). While full details of this classification system and the ways of thinking is left to Chapter 3, several general points which form a background to the work reviewed in Chapter 3 are summarised here.

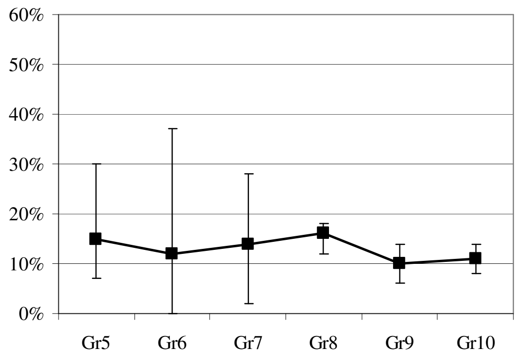
Steinle and Stacey (1998a) established that careful analysis of the response patterns of large groups of students (n=2517) reveals many variations in the ways of thinking. In order to keep the number of items on a test to a reasonable size, not all of these ways of thinking can be detected. They provide an example. *Whole number thinking* (similar to Whole Number Rule above) can be seen to consist of different variations: students who consistently choose longer decimals as larger, (known in the classification used in this thesis as *string length thinking*) and students who choose on the value of the decimal portion when considered as a whole number, (*numerator focussed thinking*). For example, although they have an otherwise similar response pattern, students using *string length thinking* will select 0.006 as greater than 0.53, but students using *numerator focussed thinking* will select correctly in this instance. Steinle and Stacey (1998a) reported that *whole number thinking* (8% of the sample of 2517 tests and labelled L1 in Chapter 3) split into *numerator focussed thinking* (156 tests, 6%) and *string length thinking* (48 tests, 2%). This paper showed that such distinctions can be made using the decimal comparison test, but also questions the usefulness and practicality of doing so.

The same paper also demonstrates that the importance of making a distinction cannot be known until data has been collected. Items were very carefully designed in order to identify a group of students who read decimals backwards (*reverse thinking*, L3, in Chapter 3). Baturo and Cooper (1995) had labelled this particular strategy "expert backwards" and a new item type (Type 6) was created to identify them by isolating items that gave different responses when read forwards and backwards. Steinle and Stacey (1998a), however, report that only 7 of the 2517 tests actually fitted the pattern. This anomaly, students expressing such thinking in interviews but not answering tests consistently enough to be detected, shows that a decimal comparison test cannot identify all misconceptions adequately – test format plays a role.

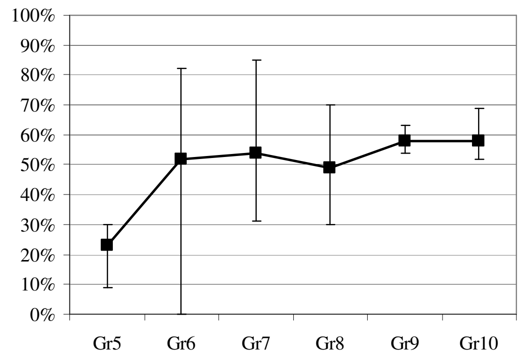
Steinle and Stacey (1998a) also demonstrated that there is considerable variation in the prevalence of the misconceptions between the schools. (These authors used the term *task expert* to describe students who completed the test with very few errors, full details in section 3.2.) For example, at one school, nearly half of the Grade 6 students exhibited L behaviour and none were *task experts* compared with another school where less than 10% of students exhibited L behaviour and over 80% were *task experts*. The overall trends (see Figure 2.2) are that L behaviour decreases with age as the prevalence of A1 (*task expert*) increases. The prevalence of S behaviour, however, is much more constant over the grades, and again there are some extreme figures. At some schools more than 1 in 3 students completing tests in Grade 6 were allocated the code S. These findings lead to the results in the present study being analysed by school groups (defined in Chapter 4) and by grade.



a) Prevalence of L by grade



b) Prevalence of S by grade



c) Prevalence of A1* by grade

Figure 2.2: Variations in prevalence of L, S and A1* by grade, Steinle & Stacey (1998a) (*A1 students are experts on the test, corresponding in spirit to ER in earlier studies)

Previously unpublished data on students' performance on test items are contained in Appendix 1. This is done so that the analysis of responses to individual items by classification, in order to reveal further aspects of decimal misconceptions, can continue.

Recently, Steinle and Stacey (2003a) reported such an analysis on the main dataset used in this thesis. Once again it shows that finer classifications are possible. In particular, it demonstrated that there are detectably different responses to the items according to whether the whole number portion is zero or non-zero and according to the length of the decimals. One of the findings of the analysis is that these two features distract students (assigned to different codes) to different degrees. The most important feature to make items easy for students exhibiting L behaviour is that the integer part is non-zero. These variations are even more accentuated in the item analysis in Steinle and Stacey (2003a) for the ways of thinking within the behaviours.

On the other hand, the most important feature to make items hard for students exhibiting A behaviour is making the number of digits after the point more than two. This analysis further points to the practical impossibility of a test accounting for all variations of thinking, and yet not knowing, before the data is collected, which features will be most important in a large sample.

Steinle and Stacey (2003a) proposed that these features provide guidance on what teaching would benefit different students. Many younger students exhibit L behaviour, so basing teaching on decimals greater than one provides an anchor for them, as there is a constant reminder of the initial "one", which was partitioned to create tenths and subsequent columns. On the other hand, older students tend to exhibit A behaviour, and for them, teaching which involves decimals with more than two digits after the point will be the most helpful.

2.6 Further evidence to support the successful diagnosis of misconceptions

This section reviews research that used interviews or additional tasks to confirm the diagnosis of student's misconceptions. It also includes research that noted the consistency of student's responses or used a statistical analysis of responses.

Kidron and Vinner (1983) tested nearly 200 students in 10th and 11th Grade (aged 15-17) with two lists of decimals to order. By asking students to write an explanation for their solutions, these researchers were amongst the first to identify S behaviour (their term, strategy 1) and note that it was "explicitly expressed by 33% of the 10th Graders and by 13% of the 11th graders", (p304).

In their first study, Nesher and Peled interviewed 21 students in Grade 6 at the end of a unit of instruction on decimals. The interviews were to probe the reasoning of the students as they completed 20 comparison items (either pairs of decimals or one decimal and a common fraction) as well as additional tasks, such as ordering 8 numbers on cards, identifying place value, number density (*how many numbers between 1.4 and 1.5?*), and hidden number comparison (*which is larger? 0._ _ _ or 0._ _*). The researchers noted that the interviews provided direct evidence that students were using particular rules (Fraction Rule, Whole Number Rule or Zero Rule) and they also note the consistency of the students' responses to the place value tasks,

> The almost complete correspondence between the pattern of each child's set of answers and the answers expected of him, when one finds he is using a certain rule, was made evident by the uniformity of the columns in Table … and confirmed the assumptions about the formulation of these rules. (p74)

In their second study, Nesher and Peled conducted a factor analysis of the 240 tests completed by students in Grades 7 to 9. This analysis confirmed the grouping of the item types and that the two major misconceptions were Whole Number Rule (Rule 1) and Fraction Rule (Rule 2),

> The results of this study supported the assumption that children are basically using two rules: Rule 1 and Rule 2. They also supported the mapping from items to rules, as it showed that all the items that were chosen to single out Rule 1 users, for example, turned out to have a high loading on the same factor. These items, so it turned out, elicited the same kind of behaviour, i.e. they "forced out" a wrong answer from Rule 1 users and a correct answer from users of Rule 2. The same phenomenon was true for Rule 2 as well. (p78)

In their small-scale study in three countries, Resnick *et al.* (1989) interviewed students to prove that there was a conceptual basis for these rules (Whole Number Rule, Zero Rule and Fraction Rule), not simply a failure of learned procedures. Interviews composed of approximately ten comparison items were used to classify the students into the rules, and then additional tasks were given to try to uncover the children's thinking. They noted the consistency of students' responses, "For the most part, the children assigned to the three incorrect rule categories gave responses perfectly consistent with the expected rules", (p16). The hidden number comparison task above (given to the USA and Israeli samples) confirmed rule use, as all 7 Expert Rule students answered correctly, *I don't know* and 18 of the 20 responses by Fraction Rule students were *shorter string*. Furthermore, they noted that inconsistencies tended to be due to specific numbers that permitted a different strategy from the standard one being used by that student (such as using knowledge that 0.5 is one half). Stacey and Steinle (1998) also noted that the item D2 (0.25 / 0.5) in Table 2.10 caused inconsistent responses by some students.

Brekke (1996) noted students' consistency between three items given in Table 2.1. For example, 86% of the 13 year old students who chose 0.649 in task 2 (corresponding to L behaviour) also chose 3.521 in task 1, and 89% of this same group chose 0.5 in task 3.

Moloney and Stacey (1997) conducted a pilot study using an 8-item decimal comparison test to classify 50 students in Grades 7 and 9. After one year, the students were retested when they were in Grades 8 and 10, this being the first longitudinal study of decimal misconceptions. While discussed below in more detail, the fact that many students gave precisely the same answers one year later provides evidence of the reliability of this decimal comparison test.

Additional evidence for the existence and correct diagnosis of student misconceptions is provided in section 3.4, when the errors made by students on the decimal comparison test are analysed.

2.7 Issues to be investigated in this thesis

Chapters 6, 7 and 8 of this thesis contain an analysis of the data with respect to the issues of persistence, hierarchy and regression, respectively. This section contains a summary of these issues arising from the literature.

2.7.1 Persistence

As reported above, many studies have used a cross-sectional methodology, which shows that misconceptions persist across the grades. Prior to the present data collection, there were few studies of how individual students progressed over time; Swan (1983) and Fuglestad (1998) used the decimal comparison test as a pre-test and post-test to determine the effectiveness of an intervention, while Moloney and Stacey (1997) used the test to determine the amount of movement by students over a year of normal teaching.

Earlier, Sackur-Grisvard and Leonard proposed that Whole Number Rule would be frequently used by students and would be *stable,* as it gives correct answers in situations with decimals of equal length, and so would not generate many conflicts for students. Actually, Fraction Rule also gives correct answers in this situation, but it was predicted to be less stable (p161).

Swan (1983) conducted research with two parallel classes. The control class was involved in “positive only” teaching, while teaching in the research class was designed to promote cognitive conflict by exposing students’ misconceptions and addressing them through class discussion and activities. The task in Table 2.5 was used as a pre-test, post-test and delayed post-test, see Table 2.11 for the distribution of students choosing each option for both classes. Swan noted (p60) that the cognitive conflict intervention was “significantly more effective at ‘permanently’ removing and correcting misconceptions”, and furthermore, that S behaviour (his term longer-is-smaller) was more difficult to remove, by a positive only approach, compared with the intervention. As no longitudinal details are provided, it is not possible to determine the movement of individuals between the tests. For example, what happened to the students who exhibited S behaviour in the pre-test? Did they persist in S, or did they move elsewhere? If students exhibited L behaviour in the pre-test and then S behaviour in the post-test, then this becomes evidence of a hierarchy, as discussed below.

Table 2.11: Distribution of students choosing various options, Swan (1983)

Choice on task		Cognitive Conflict (n=22)			Positive Only (n=25)		
		Pre	Post	Delayed	Pre	Post	Delayed
0.62	Correct	33	98	97	33	88	81
0.236	L behaviour	39	2	0	24	1	7
0.4	S behaviour	24	0	3	43	11	12

As mentioned above, Moloney and Stacey (1997) reported on the first longitudinal study of decimal misconceptions. Table 2.12 provides a cross-tabulation of the classification of each student on their 1st and 2nd tests; shaded diagonal cells indicate students who were allocated to the same rule on both tests.

Table 2.12: Cross-tabulation of rules, Moloney and Stacey (1997), p31

Rule on 1st Test	Rule on 2nd Test					Total
	ER	WNR	FR	ZR	UN	
ER	16					16
WNR		6				6
FR	2		13			15
ZR				1		1
UN	3		1		8	12
Total	21	6	14	1	8	50

The most surprising result is that very few changes were found after a period of one year. Of the 34 students who were not assigned to Expert Rule, 28 remained in the same non-expert rule on the second test. There were only six students who moved from one rule to another; five moved to Expert Rule from either Fraction Rule or Unclassified, and one student moved from Unclassified to Fraction Rule. This small longitudinal study provides evidence that the normal teaching that students receive (in this case in Grades 7 – 10) does not appear to influence students holding onto misconceptions. The issue of *persistence* will be fully investigated in Chapter 6 of this thesis.

2.7.2 Hierarchy

In an attempt to explain the various rules, Sackur-Grisvard and Leonard (1985) proposed various properties of decimals and then various steps or stages of knowledge that students may pass through. These steps were therefore considered to constitute a hierarchy: from students who ignored the decimal point completely (the most primitive) to Whole Number Rule, Fraction Rule and then Zero Rule (which they referred to as Rules 1, 2 and 3, respectively), with Expert Rule being the most advanced.

The place value tasks used by Resnick *et al.* (1989) revealed that "most" Whole Number Rule and Zero Rule "could not correctly give the value of the 5 in 1.54 and 2.45 as tenths and hundredths respectively", while Fraction Rule and Expert Rule were "mostly able to answer these questions correctly", (p21). This suggests that the Fraction Rule students are more advanced than Zero Rule, which contradicts Sackur-Grisvard and Leonard's earlier prediction.

The intervention study by Fuglestad (1998) found that, compared with the pre-test data, the post-test data showed a substantial decrease in the prevalence of L behaviour to accompany the increase in students who were able to choose correctly. What is rather surprising is that the prevalence of S behaviour, although small, actually increases; see Table 2.13. This is consistent with S behaviour being more advanced than L behaviour, as students who were initially exhibiting L behaviour may have moved to S behaviour after instruction.

Table 2.13: Prevalence of S behaviour in intervention study, Fuglestad (1998)

Sample		Pre-test	Post-test	Delayed post-test
Grade 5	(n=79)	7	12	11
Grade 6	(n=82)	7	10	8
Grade 7	(n=81)	4	7	6

Moloney and Stacey (1997) conducted a second study in which 379 students from Grade 4 to 10 were involved in a written test with two components; firstly a 15-item decimal comparison test and then 11 additional items on decimals to further probe thinking. After discarding the 31 papers that were allocated UN, the remaining 348 papers were subjected to an item analysis (provided in Moloney, 1994). Table 2.14 contains details of the additional items in the test; for easy inspection, the rows have been ranked by the overall facility. The last four columns contain the facilities for the different groups of students and are presented in a different order to earlier tables; as explained below.

Table 2.14: Facilities on additional written tasks by rule use, Moloney (1994)

Question number and details		Overall n=348	WNR n=42	ZR n=13	FR n=131	ER n=162
4	Complete 7/10 = __/100	82	81	77	76	87
11b	Which fraction is greater? 1/10 or 1/9	72	29	54	77	81
11a	Which fraction is greater? 1/3 or 1/4	71	26	38	76	81
7	Give a number anywhere between 3.8 and 3.9	64	38	38	53	83
10a	Write 0.7 as a fraction	62	24	69	53	78
5	(Number line marked 3.4 and 10 intervals to 3.5, arrow on 3.47) This number is___ ?	55	26	23	45	73
3a	How would you write the decimal for 53/100?	53	7	38	45	72
2	How would you say the number 0.53?	49	26	31	40	63
3b	How would you write the decimal for 7/1000?	45	7	38	40	60
8	Which digit in 2.875 occupies the hundredths place?	40	7	38	31	57
10b	Write 0.035 as a fraction	36	10	23	23	55
6	(Number line marked 17 and 10 intervals to 18, arrow between 17.3 and 17.4) This number is___ ?	34	5	23	21	53
9	Which decimal is equivalent to 1/5? 0.15 0.2 0.5 0.51 1.5	26	2	8	18	39

Note the general decreasing trends within each of the columns, indicating a general consensus on which of these items are the easiest/hardest. With the exception of questions 4, 8 and 10a, it can be seen that the facilities generally increase across a row, suggesting a hierarchy (lowest to highest) of Whole Number Rule, Zero Rule, Fraction Rule, and then Expert Rule. The issue of hierarchy will be carefully defined and investigated in Chapter 7 of this thesis.

2.7.3 Regression

Sackur-Grisvard and Leonard (1985) conducted a second study to determine whether the total number of decimals in the set to be ordered made a difference to the students. Nearly 300 students from Grades 7 and 8 (ages 13 and 14) were provided with sets of decimals where the target set of 2, 3 or 5 decimals were mixed among other decimal numbers to make larger sets to be ordered.

As predicted, levels of expertise dropped as the number of decimals to be ordered, and hence the cognitive load, increased. For this reason, in the present study, we have chosen to use tests consisting of multiple pair-wise decimal comparisons (referred to as *items*) rather than fewer tasks involving a list of decimals.

Furthermore, Sackur-Grisvard and Leonard noted that when students made errors, they tended to be according to the incorrect rules: Whole Number Rule, Fraction Rule and Zero Rule. The term *regression* was used to refer to the phenomenon of students being able to correctly complete the task when the set contained only 3 decimals, but not when additional numbers were included. In this thesis, *regression* will be used to describe the related but different phenomenon of a student, who answers a sequence of such items as an expert at one point in time, but is not able to do so at a later point in time, (i.e. same items at a different time instead of different items at the same time).

Swan (1983) used the term regression in this same sense (i.e. regression over time). On the task: *Ring the bigger number 56, 547, 5436* he found that all 25 students in the "positive only" class answered correctly on the pre-test. On the post-test, 2 of these students chose 56, and on the delayed post-test this had increased to 6 students. Swan concluded that the students who regressed were mechanically comparing digits from left to right as if the task was 0.5436, 0.547 and 0.56. Hence, as discussed in section 2.2, new teaching can interfere with existing ideas and then generate misconceptions.

Prior to the present study, there is little evidence within the literature, of regression in this sense, that is, regression over time. For example, Moloney and Stacey found that all 16 of the students who were allocated to Expert Rule retested the same one year later. Earlier publications from the analysis of the data reported in this thesis (e.g. Stacey & Steinle, 1999a) reported that about 1 in 10 tests coded as A were followed by a non-A test. Regression will be investigated more fully in Chapter 8.

2.8 Conclusion

This chapter has provided a brief survey of the causes of decimal misconceptions. It has reviewed the overwhelming evidence (from both large-scale written tests and small scale interviews) that patterns of errors on a carefully constructed set of decimal comparison items reveal such misconceptions. The consistent responses by students to these decimal comparison items have been well established; both from one item to another within the one test (eg Brekke, 1996; Resnick *et al.*, 1989; and Stacey & Steinle, 1998) and for one student from one test to another (Moloney & Stacey, 1997). Any errors raise the possibility that the student has a limited conceptual understanding of decimal notation. Absence of errors on such a test is a necessary but not sufficient condition for proof of understanding; to remind us of this, a student who completes a decimal comparison test with few errors is referred to as a *task expert*, see section 3.2.3 for a precise definition.

The following comment by Swan (1983) reveals the intimate connection between the test items and the misconceptions,

> It is only by asking the right, probing questions that we discover deep misconceptions, and only by knowing which misconceptions are likely do we know which questions are worth asking. (p65)

This literature review provides evidence of the evolution of a diagnostic test. The Decimal Comparison Test (DCT2) used to collect data in this study has been created from approximately two decades of research and was the most up-to-date and comprehensive test available at the start of 1997.

CHAPTER 3 THE 30-ITEM DECIMAL COMPARISON TEST

3.1 Introduction and terminology

The Decimal Comparison Test is the only source of data used in this thesis, although it draws on a range of data conducted in association with the ARC study. Of particular significance to this chapter is a series of interviews with children and adults presented in Steinle and Stacey (1998a, 2003a). These interviews revealed various ways of thinking about decimal notation, which both supported and extended those already reported in the research literature.

The Decimal Comparison Test (DCT) consists of one sheet of paper (see section 4.2.3 for the exception) on which are presented pairs of decimal numbers. The instruction at the top of the test is: *For each pair of decimal numbers, circle the one which is LARGER*. The first version of this test (DCT1) contained a list of 25 items and was used for the first 18 months of the ARC study. After an item analysis and interviews confirmed that further *item types* (defined below) were required and that other ways of thinking could be reliably identified, a second version (DCT2) containing 30 items was created and used for the remainder of the study. Appendix 1 contains copies of both tests.

The most important feature of the Decimal Comparison Test is that it is a *diagnostic* test. The choices made by a student on each item of this test create a pattern of correct and incorrect responses, which is then allocated a code. It is this pattern of responses throughout the entire test that is used to diagnose the *reason* for the choices. This test is quick to administer (an entire class can complete the test at once, taking about 5-10 minutes); easy to mark; and then the feedback to the teacher is of a similar quality to the results of a short individual interview as it refers to a student's thinking.

As this test is diagnostic, the results are not meaningfully reported as a total score or percentage. The total score on the test by a student with a particular way of thinking does not reflect "how much" the students knows; rather it reflects how many of the particular items that this student happens to get correct, are in the test. So, while two test papers may both have, say, a total of 50% correct, if the patterns of errors on the two tests are different, then the tests are allocated different codes. This has implications in determining a hierarchy in Chapter 7; for instance, it is not possible to simply rank the codes according to the total score on the test.

Terminology

The term *item* refers to a pair-wise comparison task (such as comparing 0.8 and 0.75) and a Decimal Comparison Test (DCT) is created by a sequence of such items. From this point forward items will be listed with the larger number first to be consistent with earlier publications by this author. Items are grouped into *item types*, when they elicit the same response (whether correct or incorrect) as various individuals complete the DCT. The noun *test* will be used in this thesis to indicate a student's completed script, and DCT will refer to the general test. A *code* is allocated to each test on the basis of that student's performance on various item types. As discussed in Chapter 2, rather than use the general term misconceptions, a distinction has been made between *behaviours* and *ways of thinking*. The 4 *coarse codes* (A, L, S and U) are intended to represent behaviours and the 12 *fine codes* (A1, A2, A3, L1, L2, L3, L4, S1, S3, S5, U1 and U2) are intended to represent the underlying ways of thinking.

There are two types of errors that students make that we wish to distinguish between. A student with a misconception will make *systematic* (predictable) errors as they answer the DCT. In addition, some students make *careless* errors, which are not systematic. Careless errors are answers which are incorrect and also are not in accordance with the student's own understandings. Ideally, if students make only systematic errors (i.e. no careless errors) every student's responses to every item in a given item type would be entirely consistent – i.e. all correct or all incorrect. The last section of this chapter will investigate the nature of the errors that students make on the various items on DCT2.

This chapter contains the two important components of the classification system; firstly, a description of the known ways of thinking about decimal numbers (section 3.2) which leads to the definition of the item types, and secondly, a description of the process for allocating codes, that are intended to represent these ways of thinking, to students' tests (see section 3.3). Section 3.4 provides an analysis of errors to demonstrate that the vast majority of errors by students are systematic rather than careless.

3.2 Description of ways of thinking

In Chapter 2, the behaviours of A, L, S and U were introduced. These will be used to group the ways of thinking that will be discussed in this section. The order that these behaviours are presented in this chapter (L, S, A, U) roughly corresponds to the order that they have become known. For example, Af Ekenstam (1977) considered some of the L behaviours; the S behaviours were noted in the mid 1980s; and descriptions of A and U behaviours were recorded in the late 1990s.

Firstly, however, Table 3.1 provides a useful reference for this chapter; it shows the relationship between ways of thinking and the fine and coarse codes that are allocated to students' tests. The details of the classification system, which explains the allocation of codes to completed tests, are provided in section 3.3.

Table 3.1: Matching of codes to the ways of thinking

Way of thinking	Fine Code	Coarse code
task expert	A1	
money thinking	A2	A
unclassified A	A3	
whole number thinking* & decimal point ignored thinking	L1	
zero makes small thinking & column overflow thinking	L2	L
reverse thinking	L3	
unclassified L	L4	
denominator focussed thinking & place value number line thinking	S1	
reciprocal thinking & negative thinking	S3	S
unclassified S	S4	
unclassified	U1	U
misread, misrule, mischievous,	U2	

** whole number thinking = numerator focussed thinking + string length thinking*

Before the descriptions of the ways of thinking below, note that not all ways of thinking can be allocated a separate code; they require additional items or different tasks to separate them. DCT2 does not separate all ways of thinking because space limitations on the test required some compromises to be made, and also because research conducted in parallel with the longitudinal study has revealed the existence of further ways of thinking. This issue is dealt with below and elsewhere throughout the thesis. Furthermore, the reason that S2 is not present in this table will be explained later.

3.2.1 L Behaviours

This section reviews the six ways of thinking that typically lead to longer-is-larger (L) behaviour: *decimal point ignored thinking*, general *whole number thinking* (with its subdivision into *numerator focussed thinking* and *string length thinking*)*, zero makes small thinking* and *column overflow thinking* and *reverse thinking*. Note that the fine codes intended to represent these ways of thinking start with the letter L.

Decimal Point ignored thinking:

These students completely ignore the decimal point and treat 1.34 as 134. Swan (1983) noted that an item which asked students to select the largest number from 6.78, 45.6 and 345, would detect such students as they would choose 6.78 (as 678 is larger than 546 and 345). Sackur-Grisvard and Leonard (1985) labelled this "rule of length" as such a student will typically choose the longest sequence of digits as the largest number; so 1.53 is larger than 1.8 because 153 is larger than 18. Following Resnick *et al.* (1989), the classification system of DCT2 did not separate *decimal point ignored thinking* because it was expected that it would be rarely found in the longitudinal sample.

Numerator focussed thinking:

This is one of two ways of thinking collectively referred to as *whole number thinking*. A student using *numerator focussed thinking* would produce this order for the following five decimals: 0.1, 0.04, 0.008, 0.20 then 0.034. Just as the whole number 38 is not changed by the addition of a preceding zero (038), the zeros after the decimal point are likewise disregarded. Hence, 6.3 and 6.03 would be seen as the same number: 6 wholes and 3 more parts. The disregarding of the *size* of the parts (tenths in the former and then hundredths in the latter) and concentration on the *number* of parts has suggested the name for this way of thinking. There is a body of research into students' difficulties with fractions indicating that they have trouble co-ordinating the numerator and denominator into the one number; see, for example Behr, Wachsmuth, Post and Lesh (1984) and Vance (1986). Hence, it is well established that students have difficulty ordering fractions and make a variety of errors by focussing only on the numerator or denominator in isolation– this thinking is similar in this regard.

Resnick *et al.* (1989) found that the students who were using *whole number thinking* were split into two groups when comparing 2.35 and 2.035; the students who chose 2.35=2.035 would have been using *numerator focussed thinking*.

String length thinking:

This is one of two ways of thinking collectively referred to as *whole number thinking*. These students consider the total number of digits in the decimal portion to be relevant in choosing the larger decimal. Hence, they choose 4.63 as larger than 4.8 because it is longer and likewise 4.03 as larger than 4.3. This judgement is based on the successful strategy for whole numbers, where zeros contribute both to the length and size of a number. Steinle and Stacey (1998a) report that *whole number thinking* (8% of the sample of 2517 tests) split into *numerator focussed thinking* and *string length thinking* in the ratio 3:1.

As *string length thinking* and *numerator focussed thinking* and *decimal point ignored thinking* present the same responses to the six item types on DCT2, they are not distinguished in the present study. However, supplementary items (0.53 / 0.006) and (1.3 / 0.86) were included to separate these. Almost 80% of the students allocated to *whole number thinking* chose 1.3 > 0.86, so approximately 20% may be using *decimal point ignored thinking*. (Future versions of the DCT could include at least three such items of each type to separate these three ways of thinking, but space limitations on DCT2 prevented this.)

Zero makes small thinking:

This is a slight improvement on the *numerator focussed thinking* and *string length thinking* above. The extra information these students have learnt is that a decimal with a zero or zeros in the first column(s) after the point is "small" and so they are able to make a correct decision on the comparison test for items like 4.08 and 4.5. Otherwise, they choose the longer decimal as larger. This is the thinking behind the Zero Rule (ZR) of Resnick *et al.* (1989), discussed in Chapter 2.

Column overflow thinking:

Steinle and Stacey (1998a) identified *column overflow thinking* (initially referred to as *right-hand overflow thinking*) in an interview and evidence that students use such thinking has been found in other research. For example, Hiebert, Wearne and Tabor (1991) found a student who explained that .70 > .7 as 70 tenths is more than 7 tenths (p337). This way of thinking results in the same choices on the DCT as the Zero Rule (ZR) of Resnick *et al.* (1989) but is due to different thinking.

Brekke (1996) reported that between 40% and 50% of the students aged 11, 13 and 15 involved in the KIM project wrote 0.11 for eleven tenths. Brekke commented on interviews on a similar item; *write eleven thousandths as a decimal*. He noted, (p142) that interviews confirmed that the students believed that "since thousandths are involved, they first have to write two zeros to the left of the decimal point and then the value of the [numerator]"; hence they wrote 0.0011 for eleven thousandths.

These students have created their own decimal version of column overflow. Whilst overflow to the left occurs in whole numbers, for example 120 is 12 tens, such students think 0.12 is twelve tenths. In effect, the number 12 is being squashed into the tenths column. Similarly, both 0.06 and 0.067 would be read as hundredths, first 6 then 67. Hence, a student who considers 0.11 to be eleven tenths may consider 0.11 as larger than 0.9 as 11 tenths is greater than 9 tenths.

Note that *column overflow thinking* cannot be reliably distinguished from *zero makes small thinking* by a decimal comparison test and are both allocated the code L2; an interview could be used to separate such students.

Reverse thinking:

A student who has not heard the "*th*" in the place value names (or who has heard it but disregards it as it has no meaning for them) may believe that the decimal columns represent more whole numbers but written in the reverse order: i.e., (point) tens, hundreds, thousands, etc. Steinle and Stacey (1998a) found evidence for this confusion in an interview with a Grade 5 student. When asked to read 0.163 from a card she replied,

> ..one hundred and sixty three....because when we do that in class we had a tens column and a hundreds column and a thousands column....I'm not sure if its just one hundred and sixty three or its 1 ten, 6 hundreds and 3 thousands. (p549)

Other researchers have noted students who appear to be using *reverse thinking*. For example, Baturo and Cooper (1995) found several students who used this strategy to compare decimals, which they labelled as EB (expert backwards). Interestingly, these 3 students voiced this strategy in interviews but did not use it consistently on all items in the written test. Similarly, Behr and Post (1988) discuss an interview with a student in which 0.37 and 0.73 were to be compared. The student stated that 0.37 > 0.73 because, "0.37 was 3 tens and 7 hundreds (not 3 *tenths* and 7 *hundredths*) and 0.73 was 7 tens and 3 hundreds (not 7 *tenths* and 3 *hundredths*)", (p223).

During interviews, Resnick *et al.* (1989) noted that some students who had been classified as using an L behaviour, "reversed the order of the digits 6 (tenths) and 2 (hundredths), writing 0.026 or 2.6, which would be the correct ordering if the number were 6 tens and 2 hundreds" (p21). Furthermore, Irwin (2001) noted that some students (p402) read 0.01 as *one hundred*. These errors may appear to be just a "slip of the tongue", but for some students this may reflect a deeply held misconception. These reports lead us to expect that this way of thinking is common since it is frequently encountered when interviewing students. However, it will be seen later that it is rarely encountered in responses to the DCT.

This concludes the discussion of the ways of thinking that lead to L behaviour, reported within the literature. Before proceeding to a discussion of the ways of thinking that lead to S behaviour, Figure 3.1 (Steinle, Stacey and Chambers, 2002) provides an illustration of how students with three different ways of thinking might order various decimals. In sample 1 (an example of L behaviour), 1-digit decimals are interpreted as smaller than the 2-digit decimals; while in both samples 2 and 3 (examples of S behaviour) the opposite order applies (i.e. 1-digit decimals are larger than 2-digit decimals). Note how the sets of decimals of various lengths are not interspersed, rather they are grouped together.

	Sample 1: L behaviour *whole number thinking*	Sample 2: S behaviour *denominator focussed thinking*	Sample 3: S behaviour *reciprocal thinking*
	2	2	2
	.	1.9	1.1
	.	1.8	1.2
	1.6792	1.7	1.3
	.	1.6	1.4
	.	1.5	1.5
	1.999	1.4	1.6
	1.998	1.3	1.7
	.	1.2	1.8
	.	1.1	1.9
	1.101	1.99	1.10
	1.100	1.98	1.11
	1.99	.	1.12
	1.98	.	.
	.	1.51	.
	.	1.50	1.49
	1.51	1.49	1.50
	1.50	.	1.51
	1.49	.	.
	.	1.12	.
	.	1.11	1.98
	1.12	1.10	1.99
	1.11	.	1.100
	1.10	.	1.101
	1.9	1.999	.
	1.8	1.998	.
	1.7	.	1.998
	1.6	.	1.999
	1.5	1.101	.
	1.4	1.100	.
	1.3	.	1.6792
	1.2	1.6792	.
	1.1	.	.
	1	1	1

Figure 3.1: Three samples of ordering decimals (Steinle, Stacey & Chambers, 2002)

3.2.2 S Behaviours

This section reviews the four ways of thinking that typically lead to shorter-is-larger (S) behaviour: *denominator focussed thinking*, *place value number line thinking, reciprocal thinking* and *negative thinking*. Note that the fine codes intended to represent these ways of thinking start with the letter S.

Denominator focussed thinking:

Previous researchers have identified these students; for example, Fraction Rule (FR) reported by Resnick *et al.,* (1989). Putt (1995) found this thinking in interviews with pre-service teachers, "Firstly I changed all of the decimal fractions into common fractions. Then I looked to find the fractions with the denominator of 10,000 as this will be the smallest group of fractions", (p9).

Based on the fact that *one* tenth is larger than *one* hundredth, they incorrectly generalise to: *any number* of tenths is larger than *any number* of hundredths. While they may or may not have an image of the equivalent fraction, the consideration of the size of the parts (in isolation) has determined their choice and is reflected in the name of this group.

As these students are likely to demonstrate reasonable understanding of place value names and make sensible decisions with decimals of equal length, (see Sample 2 in Figure 3.1) it would be easy for their wrong thinking to go undetected in the classroom.

Place Value Number line thinking:

These students make similar responses to students with *denominator focussed thinking* on the Decimal Comparison Test, but the explanation differs. Stacey, Helme and Steinle (2001) present details of an interview with 'Stuart', a pre-service teacher who, when asked to draw a number line, first wrote the numbers 10, 0 and 0.1 (from left to right). Figure 3.2 demonstrates how a student might consider the place value columns to represent a pseudo number line. The meaning of the place value column (e.g. hundreds) becomes conflated with the *set of numbers in the hundreds*, creating a *place value number line*. Note that this incorrect number line contains *correct* information about the whole numbers, in particular, that the set of 4-digit whole numbers are separated from, and larger than, the set of 3-digit whole numbers. Likewise, the set of 3-digit whole numbers are separated from, and larger than, the set of 2-digit whole numbers. So, this way of thinking provides correct answers in the limited context of whole numbers. Note that the reversed direction of this *place value number line* may be a useful clue to diagnose such thinking.

Putt (1995) found this thinking in interviews with pre-service teachers, "I put 0.6 as the largest because I figured it was closest to having a whole number. That's why I put 0.66 smaller because I thought that having a hundredth in the hundredths column made it smaller", (p9). It has also been noted by Irwin (1996), who commented, "other orderings combined some understanding of decimal fractions with number line and place value columns", (p249).

A student with *place value number line thinking* is likely to consider that 0.6 is less than zero, because zero is in the ones column and 0.6 is in the tenths, which is on the right hand side of the ones column, hence smaller. Later versions of the decimal comparison test following DCT2 include such items to distinguish *place value number line thinking* from *denominator focussed thinking*.

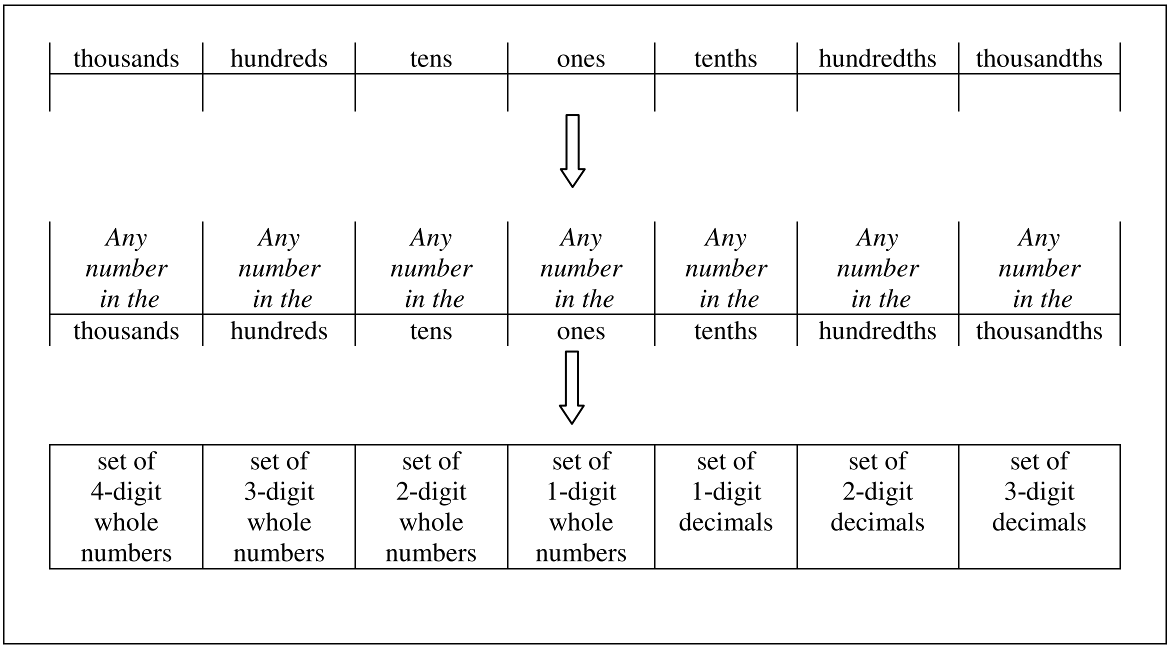


Figure 3.2: The origins of a Place Value Number Line

Reciprocal thinking:

As noted in Chapter 2, several researchers, for example, Markovits and Sowder (1991) and Swan (1983), have identified students who are tempted to consider the decimal portion of the decimal number as analogous to the denominator of a fraction. For example, they see the fraction *a/b* as somewhat similar to the decimal *a.b*, the logical consequence of which is to consider 3.87 as smaller than 3.86 (as larger denominators make smaller fractions). Brekke (1996) reported that more than one in three Grade 6 students chose the distractor 8.20 when shown a diagram illustrating 8 parts shaded out of 20. Similarly, Irwin (2001) reported students who wrote one hundredth (1/100) as 0.100 and Moss and Case (2002) reported a student that said "because one eighth is probably the same as decimal eight", (p149).

These students have attempted to connect decimals with fractions but have difficulty, as the decimal form does not explicitly indicate the denominator. Hence, when faced with the notation 0.3 they perceive that there are 3 parts rather than 10 equal parts of which 3 are chosen. See Sample 3 in Figure 3.1 for how these students might order decimals between 1 and 2.

Nor is this thinking limited to school students. Steinle and Stacey (1998a) found evidence for this particular confusion between decimals and reciprocals in an interview with a tertiary student. As noted in section 2.2, Stacey, Helme and Steinle (2001) proposed that this way of thinking can be explained by the mirror metaphor.

Negative thinking:

These students are confusing decimals with negative numbers. Steinle and Stacey (1998a) referred to an interview with a pre-service teacher 'Amanda' who chose 0.20 > 0.35, 2.516 > 2.8325 and 7.63 > 7.942. While such choices are consistent with *reciprocal thinking* discussed above, Amanda's explanation was different,

> I was thinking along a number line and considering decimal numbers to be equivalent to negative numbers. Therefore -20 was larger than -35. …I felt more comfortable selecting the number with the least digits as I thought the longer the number, the further it was down the number line in the negative direction. (p550)

Putt (1995) also found pre-service teachers who expressed this view for decimals less than one, "Looking back, I think that on a number line on the negative side the farther from zero (is the) smallest and closer to zero is the biggest", (p11).

Stacey, Helme, Steinle, Baturo, Irwin and Bana (2001) reported a pre-service teacher who had written the following comment regarding a comparison of 0 with 0.6,

> Comparing 0 with a decimal may also lead to confusion as to which is the smallest, as 0 represents 'nothing' to young students and it may be difficult for them to understand that something can be 'smaller' than this. (p215)

The cause of this way of thinking may be confusion with another notation (a dash or a dot) or it could also be argued that it was due to interference of new teaching. We predict that the prevalence of this way of thinking would be higher in Grade 7 or 8 when negative numbers are introduced to students. In addition, the use of scientific notation in secondary school (0.0003 can be written as 3×10^{-4}) can also remind students that decimals are somehow related to negative numbers. As noted in section 2.2, Stacey, Helme and Steinle (2001) proposed that this way of thinking can be explained by the mirror metaphor.

Note that *reciprocal thinking* and *negative thinking* are unable to be separated on DCT2 and so both are allocated the code S3. (Additional items such as comparing 0.6 with 0 would separate these, as we predict that *reciprocal thinking* would choose correctly, and *negative thinking* would choose incorrectly).

3.2.3 A Behaviours

Much of the research on decimal misconceptions has focussed on students who were unable to choose correctly on an item such as Swan's item, (comparison of 0.236, 0.4, 0.62) which indicated that they were choosing according to the relative number of digits within the numbers (i.e. L or S behaviours). After investigating the thinking behind the L and S behaviours, it is now appropriate to consider the students who would choose correctly on the above task. While some of these students will have a rich conceptual understanding of all aspects of decimal notation, others will be shown to be following *procedures* for comparing decimals and have little conceptual understanding.

Students exhibiting A behaviours correctly compare "straightforward" pairs of decimals. The *task experts* have high scores on all item types and are discussed below, as are students using *money thinking*. A full investigation and discussion of A behaviours is provided in Chapter 8.

Task experts

Some students who answer all item types correctly are indeed experts, having a solid understanding of decimal notation; others may not have a good understanding, but correctly follow either of the expert algorithms to order decimals. They may use the *annexe zero algorithm* or they may use the *left-to-right digit comparison algorithm*. Both of these algorithms can be used without any understanding of why they work.

Resnick *et al*. (1989) had noted that students being classified as experts may not understand the conceptual basis for decimal comparisons, and may have arrived at their correct answers "on the basis of purely surface and syntactic considerations", (p25). Moloney (1994) found that only 40% of the 162 students being classified as experts on decimal comparison items were able to choose 0.2 as the decimal equivalent of 1/5 in a multiple choice item (see Table 2.14). Furthermore, less than 60% could identify which digit occupies the hundredths place in 2.875. Stacey and Steinle (1998) interviewed students who had made very few errors on DCT1. Some students were unable to answer questions such as 'write a number between 0.35 and 0.36' or to draw numbers on a number line.

Hence, Steinle and Stacey (1998a) introduced the term *task expert* to refer to students who made very few errors on the task of decimal comparison, to emphasize that this was evidence that a student could complete the comparison task and they may not have a wider knowledge of decimals.

Money thinking:

Stacey and Steinle (1998) referred to tertiary students 'Nancy' and 'Emily' as evidence that some students may truncate or round to two decimal places when comparing decimals. (Furthermore, Steinle and Stacey (2003a) suggest that some students will consider just the first digit after the point; effectively truncating numbers to one decimal place).

Nancy used the context of money (hence the term *money thinking*) to compare decimals. She explained that 4.08 was smaller than 4.7 as $4.08 was less than $4.70, but was unsure what to do with the pair 4.4502 and 4.45,

> When the numbers are the same in the same spot I get very confused….Does the number get bigger or smaller with more numbers on the end?.. When the number after

> the decimal point is different the question is easier, but when they are the same, I don't know what rule to apply. (p60)

Nancy is proof that an over-reliance on the context of money does not assist her to make sense of decimals with more than two places, as predicted by Brekke (1996). Emily was able to correctly compare decimals using the context of percentages. In effect, Emily was multiplying the decimal by one hundred and then ignoring any remaining decimal portion. This strategy does not provide a clear choice when comparing say 0.4502 and 0.45 as both are 45%.

Note that as well as money and percentages, students using (exclusively) the context of metres and centimetres will also experience the same problems as Nancy and Emily. The procedure of constantly rounding answers to two decimal places in secondary school may reinforce these procedures or limited contexts.

In an attempt to determine whether students believed that 8.41242 and 8.41 were equal, a new version of the decimal comparison test was created (see Steinle & Stacey, 2001) with the instruction: *Put a ring around the bigger decimal OR write = between them.* This test (DCT0) was used by a sample of approximately 300 students in total: 100 pre-service elementary teachers from one Australian university, and 200 school students from 3 schools in Japan. The test contained seven items such as (8.41242 / 8.14) and (0.73222 / 0.73). Of the students who made errors on these items, about one in four chose the equals option, and three in four chose the wrong inequality.

3.2.4 U Behaviours

For the majority of students, their pattern of correct and incorrect responses on the DCT matches the predicted responses from one of the above known ways of thinking; those tests that do not are referred to as *unclassified.* It may be the case that the student has an unknown way of thinking that they use consistently, or they have a mixture of several of these ideas, which they use inconsistently. There is some evidence (Moloney & Stacey, 1997) that students who are beginning to develop more sophisticated ideas about decimals will answer most inconsistently and this is a hypothesis that the present study will explore.

Misread / Misrule / Mischievous:

Amongst the tests that were unclassified were a group of tests with almost every item wrong. Three possibilities are presented for explaining this phenomenon; either a *task expert* who misreads the instructions, circling the *smaller* number throughout the test, OR a *task expert* who is intentionally choosing the incorrect answer (being mischievous), OR a student following an algorithm to compare decimals (such as *annexe zero algorithm* or the *left-to-right digit comparison algorithm*) but then believing that there is a reversal in size (by loose analogy with fractions and negative numbers). It is hoped that following such students longitudinally in the present study will shed light on the reasons for their choices.

3.3 Allocation of codes to tests

As explained earlier, a code is allocated to a test, not based on the total score, but on the performance on various item types. In this section, the allocation of coarse and fine codes is explained with reference to DCT2 and the variations for DCT1 are then provided. (See Appendix 1 for copies of both tests.)

Decimal comparison items are grouped into item types so that students with different ways of thinking can be distinguished by their patterns of answers. The criteria for each item type need to be carefully designed so that this is possible. For example, in order to identify *reverse thinking*, the items used in every item type need to give consistent responses when read "forwards" (i.e. normally) and also consistent responses when read "backwards", which reverse thinkers will do. Construction of a good decimal comparison test is therefore a complicated task. The definitions of the item types are given in Appendix 1.

3.3.1 Allocation of coarse codes (DCT2)

Firstly, a student's test is assigned one of 4 codes (A, L, S or U) intended to represent the behaviours discussed in Chapter 2. A coarse code is allocated to a student's test according to the pattern of correct and incorrect responses to items within Types 1 and 2 as in Table 3.2. For consistency with earlier publications by this author, the correct (larger) decimal is listed first, but obviously the order is varied on the test.

Type 1 items are designed to be answered *correctly* by students with S behaviour, while students with L behaviour choose *incorrectly*. (See Appendix 1 for definitions of the item types.) The opposite responses will be noted for the Type 2 items. Students who answer correctly on both Types 1 and 2 are allocated the code A and students whose tests do not meet these criteria are coded as U. Note the allowance of one deviation per item type; hence a student with 4 correct in Type 1 and 4 correct in Type 2 will still be coded as A. Likewise, a student with 1 correct in Type 1 and 4 correct in Type 2 is still coded as L. (Section 3.4 provides an analysis of students' errors which demonstrates that these coarse codes are indeed meaningful.)

Table 3.2: Expected responses by students with particular behaviours

Details of items, (larger number first)			Coarse codes			
			A*	L*	S*	U
Type 1	4.8	4.63	√	X	√	none of A, L or S
	0.5	0.36	√	X	√	
	0.8	0.75	√	X	√	
	0.37	0.216	√	X	√	
	3.92	3.4813	√	X	√	
Type 2	5.736	5.62	√	√	X	
	0.75	0.5	√	√	X	
	0.426	0.3	√	√	X	
	2.8325	2.516	√	√	X	
	7.942	7.63	√	√	X	

** One deviation (per type) from the expected responses is allowed.*

3.3.2 Allocation of fine codes (DCT2)

After allocation of the four coarse codes (A, L, S and U), consideration of additional items from Types 3 to 6 allows each coarse code to be subdivided to create a total of 12 *fine* codes. For example, a student's response to Type 3 items (e.g. 4.7 / 4.08) will be used to separate some of the students with L behaviour; *zero makes small/column overflow thinking* will choose correctly while *whole number thinking* will choose incorrectly. Students who have been allocated to A behaviour may be split on their responses to the Type 4 items (e.g. 4.4502 / 4.45) with students who use *money thinking* making errors. Students who have been allocated to S behaviour may be split on their responses to the Type 5 items (e.g. 0.4 / 0.3) with students who use either *reciprocal thinking* or *negative thinking* making errors while students using *denominator focussed* *thinking* or *place value number line thinking* would be expected to choose correctly. The Type 6 items (0.42 / 0.35) were designed to separate students with *reverse thinking* (choose incorrectly) from *whole number thinking* (choose correctly).

Rather than list every item as in Table 3.2 above, only the item types are provided in Table 3.3 and the performance is indicated by *Lo* (indicating Low performance, i.e. at most one *correct* answer) and *Hi* (indicating High performance, i.e. at most one *incorrect* answer). The classification scheme in Table 3.3 includes the Type 1 and 2 items for completeness; note that the three codes A1, A2 and A3 have the same response to the *core items* (Types 1 and 2); likewise L1, L2, L3 and L4, etc. The actual items in DCT2 are given in Appendix 1, along with the definitions of the item types. The absence of the codes S2 and S4 is explained in the next section.

Table 3.3: Classification scheme for allocating fine codes to DCT2 tests

Item Type (number)			Fine Codes											
			A1	A2	A3	L1	L2	L3	L4	S1	S3	S5	U2	U1
Core	1	(5)	Hi	Hi	Hi	Lo	Lo	Lo	Lo	Hi	Hi	Hi	Else and few* correct	Else
	2	(5)	Hi	Hi	Hi	Hi	Hi	Hi	Hi	Lo	Lo	Lo		
Non-Core	3	(4)	Hi	Hi	Else	Lo	Hi	Lo	Else	Hi	Hi	Else		
	4	(4)	Hi	Lo		Hi	Hi	Hi		Lo	Lo			
	5	(3)	Hi	Hi		Hi	Hi	Hi		Hi	Lo			
	6	(3)	Hi	Hi		Hi	Hi	Lo		Hi	Lo			

Hi=High (at most one error in the set of items for that type)
Lo=Low (at most one item correct in set)
** up to 6 correct*

If a test is coded as A, but it does not fit the predictions for A1 and A2, it is then coded as A3, rather than U. A similar situation holds for L4 within L and S5 within S. Hence A3, L5 and S5 are allocated to tests with consistent patterns on Types 1 and 2 (core items), but inconsistent patterns on Types 3 to 6. These three codes are referred to as *fine unclassifieds.*

Initially, the code U2 was not included in the formal classification; rather it was used to indicate to teachers that perhaps the student was a *task expert* who had misread the instructions (as the smaller rather than larger decimal in each item was chosen fairly consistently). The definition of U2 was not as stringent as other codes; it refers to tests that were unclassified and also had a small total score (up to approximately 6 correct).

Apart from the code U2, note the strictness of this classification system; Type 1 contains 5 items, so scores of 0 or 1 are referred to as Low (*Lo*), and scores of 4 or 5 are referred to as High (*Hi*). A test with a score of 2 or 3 is automatically coded as U. Table 3.4 contains the number of items in each item type, the corresponding questions on DCT2, and the scores that correspond to Low and High performance.

Table 3.4: Details of item types within DCT2 (number of items and performance)

Item types	Number of items	DCT2 questions*	Performance	
			Low	High
1	5	6, 7, 8, 9, 10	0, 1	4, 5
2	5	16, 17, 18, 19, 20	0, 1	4, 5
3	4	12, 13, 14, 15	0, 1	3, 4
4	4	21, 22, 23, 24	0, 1	3, 4
5	3	3, 4, 5	0, 1	2, 3
6	3	26, 27, 28	0, 1	2, 3

* *six supplementary items are not listed here, see Appendix 1 for full test*

3.3.3 Variations on the DCT1

In 1995 and 1996, the 25-item DCT1 was used to collect data and tests were allocated to the rules WNR, FR, ZR, ER and UN, as used by Resnick *et al.* (1989), and discussed in Chapter 2. As mentioned earlier, the careful analysis of this data (plus interviews) revealed that an improved classification system could be constructed and DCT2 was created with 16 common items to DCT1. The shift from DCT1 to DCT2 was to enable more detailed diagnosis, but in order to maximise the longitudinal data, it was desirable to link performance on the two tests.

Two tables (similar to those presented earlier) will now be provided for the allocation of codes to DCT1 tests; Table 3.5 provides details of the classification system for the DCT1 tests, and Table 3.6 provides the corresponding details of the performance on each item type.

Table 3.5: Classification scheme for allocating fine codes to DCT1 tests

<table>
<tr><th colspan="3" rowspan="2">Item Type (number)</th><th colspan="12">Fine Codes</th></tr>
<tr><th>A1</th><th>A2</th><th>A3</th><th>L1</th><th>L2</th><th>L4</th><th>S1</th><th>S2</th><th>S3</th><th>S4</th><th>U2</th><th>U1</th></tr>
<tr><td rowspan="2">Core</td><td>1</td><td>(4)</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Lo</td><td>Lo</td><td>Lo</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Hi</td><td rowspan="5">Else and few correct</td><td rowspan="5">Else</td></tr>
<tr><td>2</td><td>(3)</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Lo</td><td>Lo</td><td>Lo</td><td>Lo</td></tr>
<tr><td rowspan="3">Non-Core</td><td>3</td><td>(4)</td><td>Hi</td><td>Hi</td><td rowspan="3">Else</td><td>Lo</td><td>Hi</td><td rowspan="3">Else</td><td>Hi</td><td>Hi</td><td>Hi</td><td rowspan="3">Else</td></tr>
<tr><td>4</td><td>(3)</td><td>Hi</td><td>Lo</td><td>Hi</td><td>Hi</td><td>Lo</td><td>Lo</td><td>Lo</td></tr>
<tr><td>5</td><td>(2)</td><td>Hi</td><td>Hi</td><td>Hi</td><td>Hi</td><td>2</td><td>1</td><td>0</td></tr>
</table>

Table 3.6: Details of item types within DCT1 (number of items and performance)

Item types	Number of items	DCT1 questions	Performance	
			Low	High
1	4	1, 4, 10, 13	0, 1	3, 4
2	3	6, 15, 17	0, 1	2, 3
3	4	2, 5, 8, 14	0, 1	3, 4
4	3	3, 9, 12	0, 1	2, 3
5	2	21, 22	0	1, 2

Note that the codes S2 and S4 were not used earlier; they are now explained. When two of the supplementary items in DCT1 were analysed (0.4 / 0.3 and 1.85 / 1.84) it was found that the Fraction Rule students did not treat these items consistently; see Table 2.10. In particular, there were considerable numbers of Fraction Rule students choosing incorrectly on both items, which was not predicted. (In fact, DCT1 was based on Resnick *et al.* (1989), which assumed that Fraction Rule students were using *denominator focussed thinking*, whereas many are in fact using *reciprocal thinking*.)

Hence, these two items were used to create a new item type (Type 5) and the three possible scores of 2, 1 and 0 were then used to allocate the three codes of S1, S2 and S3, respectively, as in Table 3.5. At this point S4 was used to indicate S but not consistent on Types 3 to 5. The students with tests allocated S2 may well be S1 or S3 students who made one careless choice. It was hoped that the longitudinal analysis would shed more light on the S2 students; as explained in the next chapter, this was not the case.

Note also that L3 (*reverse thinking*) is unable to be distinguished from L1 (*whole number thinking*), as DCT1 contained no Type 6 items. In fact, the low prevalence of

L3 from DCT2, less than 1%, has led to L3 tests being combined with L4 tests in this thesis and confidence that the L1 allocation in DCT1 is almost always correct.

A cross-tabulation of the old rules and the “best” available new classification (given the items within DCT1) for these 1852 tests is provided in Table 3.7. It is clear from this table that almost every ER test was allocated to A1; similarly for Whole Number Rule to L1 and Zero Rule to L2. The two groups that were involved in reclassification were the Fraction Rule tests (split into S1, S2, S3 and A2) and the Unclassified tests which were allocated to many different new codes.

Table 3.7: Imposing new classification on DCT1 tests (n=1852*)

Old	New code												Total
rule	A1	A2	A3	L1	L2	L4	S1	S2	S3	S4	U1	U2	
ER	560									3			563
FR		41					98	60	96				295
WNR				453		2					2		457
ZR					65	1							66
UN	33	67	34	9	13	99	13	8	6	38	148	3	471
Total	593	108	34	462	78	102	111	68	102	41	150	3	1852

**3 dropped later as repeating students don’t fit into cohorts, so later n=1849*

3.4 The nature of errors

The following discussion relates to DCT2 and is provided to verify that the vast majority of errors made by students are systematic rather than careless. The data comes from testing which took place in 1997. Of the 3531 tests completed, exactly 1200 tests had every answer correct and were discarded from the following analysis, leaving 2331 tests with at least one error. (Note that the inclusion of these tests with no errors would simply increase by 1200 the number of students in the following discussion with full score on each item type.)

The distribution of the scores on each item type is provided in Figure 3.3. Both Type 1 and Type 2 contain 5 items and a student might score 0, 1, 2, 3, 4, or 5 out of 5. The remaining items types have fewer items and hence lower maximum scores (4 for Types 3 and 4, and 3 for Types 5 and 6).

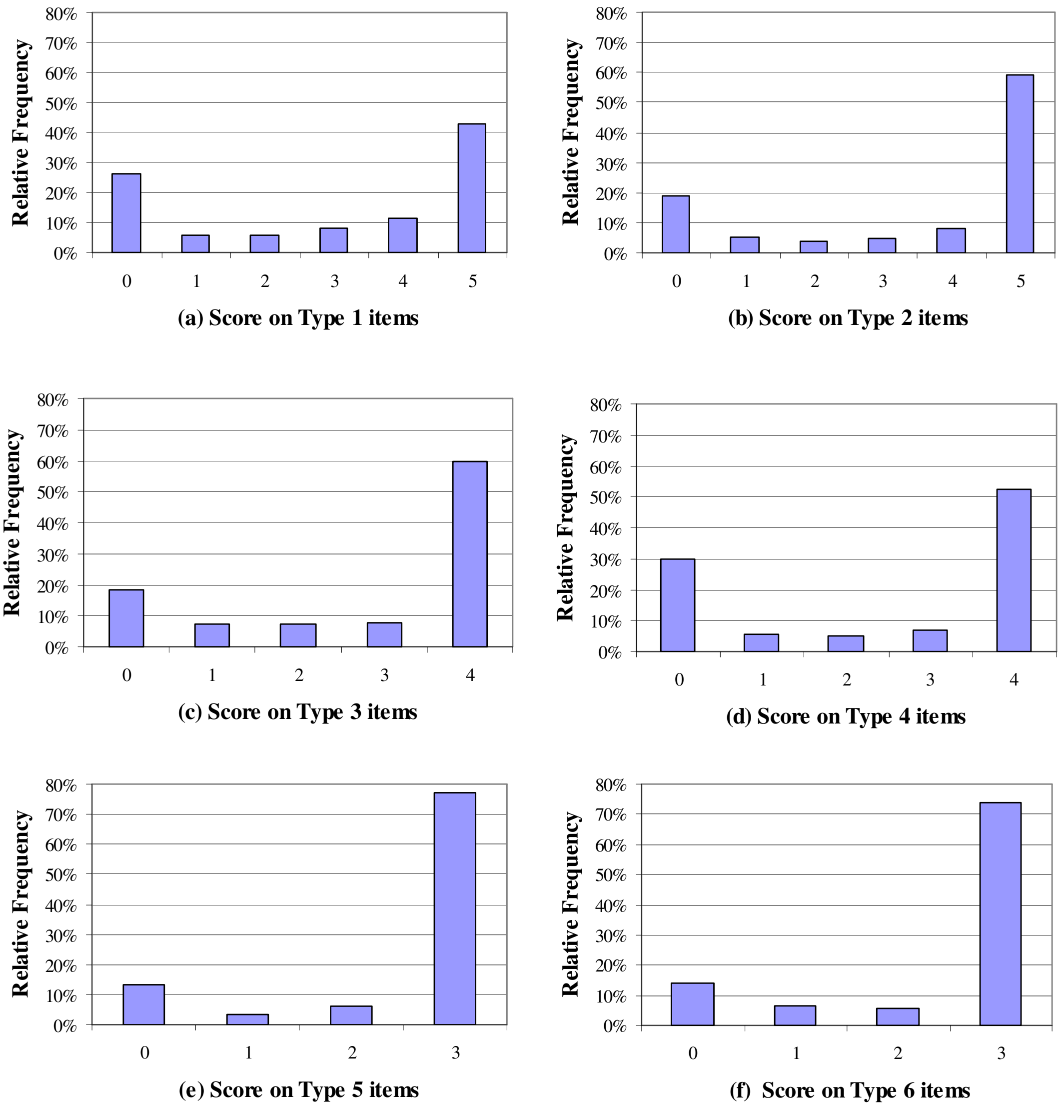


Figure 3.3: Scores on Types 1 to 6, Steinle and Stacey (2003a), p636

It is clear from these six graphs in Figure 3.3 that in any particular item type, more students score full marks than any other score and the next most popular score is zero. This leads to the conclusion that, in general, the errors made by students are not careless errors, rather they are systematic and predictable. For example, once a student is observed to make an error on the first item within, say, the Type 1 items, it is very likely that this student will then make four more errors on the remaining four such items.

In another study with a different type of test, it might be expected that the left-most columns in Figure 3.3 are due to some "low ability" students who make many errors. It will now be shown that this is not the case by considering the scores on Type 1 items and Type 2 items simultaneously. There are 36 possible scores on Type 1 and Type 2 items (0 to 5, in both types). Reference to Table 3.2 provides the following definitions of the coarse codes where (X,Y) indicates a score of X for Type 1 and a score of Y for Type 2:

- A is allocated to (5,5), (4,5), (5,4) and (4,4)
- L is allocated to (0,5), (0,4), (1,5) and (1,4)
- S is allocated to (5,0), (4,0), (5,1) and (4,1)
- U is allocated to the remaining 24 possibilities.

The resulting distribution from the above-mentioned 2331 tests is provided in Figure 3.4. (Note this differs from Figure 2 in Steinle and Stacey (2003b) which was based on all 3531 tests completed in 1997.) There are, in fact, very few students who score 0 or 1 for both types (low columns at the front corner); rather the two corners come from scores of (0,5) and (5,0). There are three clear peaks and the following points can be concluded:

- Not surprisingly, a considerable number of students answer all 10 items correctly. The column at the back corner (5,5) in Figure 3.4 would be considerably taller than all other columns if the additional 1200 tests with full score were included in this sample.
- The very tall columns at the side corners, corresponding to L and S behaviours, confirm the validity of the L and S constructs. (In fact, if only the 1867 papers with one or more errors in these 10 items are considered, then 59% are coded as either L or S).
- The very low columns in the centre of the graph demonstrate again that students, on the whole, are not making careless errors on this test.

Hence, while Figure 3.3 indicated that the sets of Type 1 and Type 2 items are reasonably internally homogeneous, Figure 3.4 now shows that these item types are different to each other and hence that the classifications to L, S and A from performance on Types 1 and 2 are indeed meaningful.

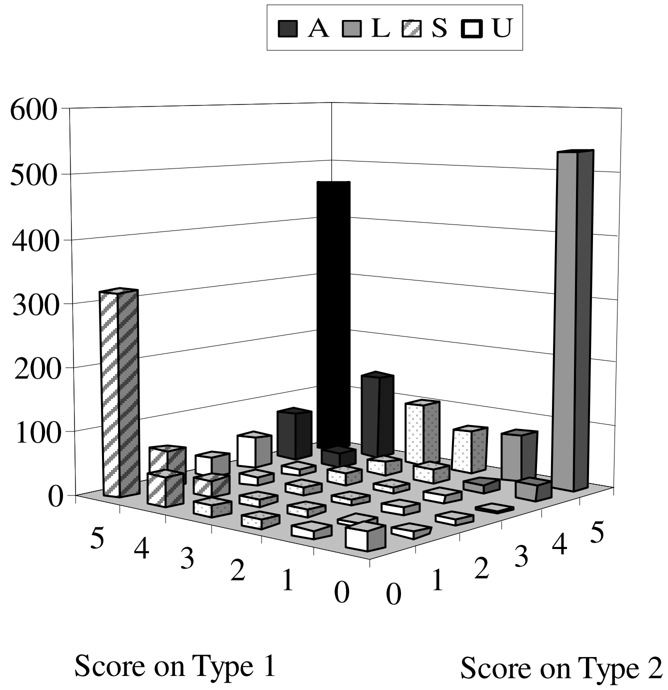


Figure 3.4: Actual distribution of scores on Types 1 and 2 based on data from 1997 (2331 tests with a maximum score of 29/30)

Figure 3.5 is provided as a contrast; this would be the distribution of scores if students answered randomly on each item with a probability of being correct on any item of 0.5 and answered the items independently. The majority of students would, in this case, have scores of 2 or 3 on both Types 1 and 2. A choice other than one half for the probability of being correct or unequal probabilities for Type 1 and Type 2 will shift the peak of probabilities from the centre, but not change the general single peaked shape.

Details of the item-by-item analysis of DCT1 are contained in Stacey and Steinle (1998) and a similar analysis of DCT2 items is provided in Steinle and Stacey (2003a). Such fine-detailed analyses have provided additional information about students' responses within the item types.

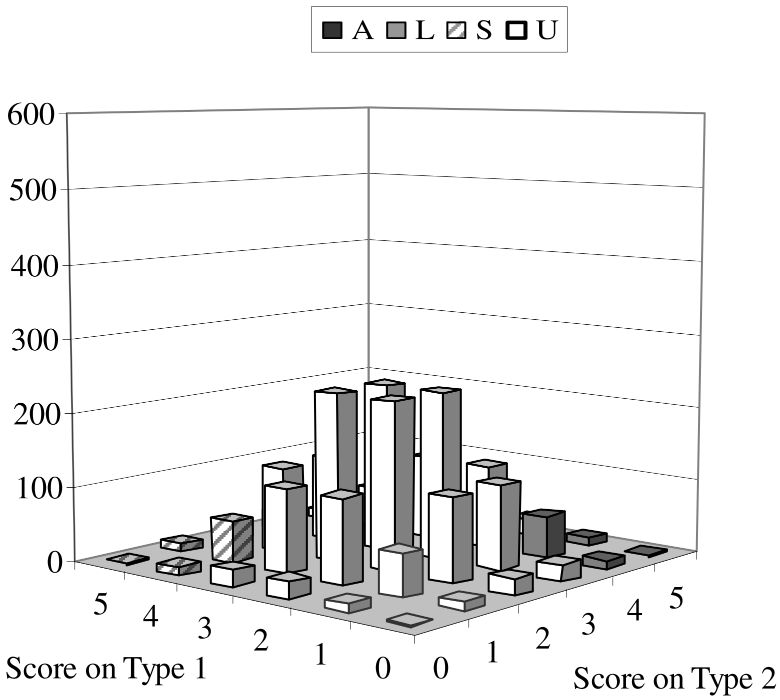


Figure 3.5: Expected frequency distribution for scores on Types 1 and 2 assuming probability of correct on any item is 0.5 and answers selected independently

3.5 Conclusion

The two versions of the Decimal Comparison Test (DCT1 and DCT2) that were used to collect data in the ARC study have been discussed in this chapter, as well as the procedures for allocating codes to completed tests. The relationship between the codes (both coarse and fine) and the ways of thinking has been established, and each of the ways of thinking has been described, with references to the literature.

Furthermore, evidence has been provided to indicate that, on the whole, students do not make careless errors when completing these tests. Rather, students make systematic errors that reveal their misconceptions. The items within each item type have been shown to be adequately internally homogeneous and the Type 1 and Type 2 items have been shown to be different from each other and hence effectively separate students with either L behaviour or S behaviour from students with A behaviour.

The unusual nature of this diagnostic test has been described in this chapter. For example, two students may have the same number of correct answers on their tests but be allocated different codes as their pattern of correct and incorrect responses differ. This has implications, for example, in the determination of a hierarchy (see Chapter 7).

CHAPTER 4 METHODOLOGY

4.1 Introduction and terminology

The ARC study started data collection in the second semester of 1995 and testing was conducted twice each year until the middle of 1999. Students from Grades 4 to 10 were involved in testing and, as this was a longitudinal study, it was the intention to repeatedly test the same students wherever possible. This data collection from the ARC study has generated approximately 11000 tests; the longitudinal analysis of these tests being the focus of this thesis.

This chapter has several purposes. It provides details of the data collection in the ARC study, and the procedures used in the data analysis, such as the tracking of students. It also provides details of some features of the data that set the scene for the chapters to follow.

Section 4.2 contains details of the procedures used in the data collection in the ARC study. The tests, the sample, the testing schedule, and the procedures for conducting the testing are all described, as well as some of the consequences for the later analysis.

Section 4.3 contains details of the procedures used in the data analysis, such as how students were tracked through the longitudinal testing as they moved from grade to grade and then from school to school. The tension between focussing on individual *tests* or individual *students* is discussed as many issues can be analysed from both points of view.

Section 4.4 contains meta-data, that is, tables which illustrate details of the testing. For example, the number of tests completed by students in a given grade; the number of students who completed, say, 3 or more tests; the grade of a student when they joined the study; the time between the tests, etc.

Terminology

School level refers to either primary school (up to Grade 6) or secondary school (from Grade 7). In order to follow students from primary school to secondary school, schools in the same geographical area were grouped; for example *School Group A* (SGA) consists of one secondary school and two of its “feeder” primary schools. Table 4.2 will provide precise details of the school groups.

Testing was generally conducted twice a year over several years. The timing of the testing is referred to in two ways; *Semester 1* and *Semester 2* refer to the first and second half of the year, while *Testing Periods 1* to *8* refer to the eight occasions over the four years that testing took place. Testing Period 1 refers to Semester 2 of 1995, and Testing Period 2 refers to Semester 1 of 1996, etc.

A *cohort* refers to a group of students who are in the same grade in the same calendar year and are expected to move through the grades together; for example, students in Grade 6 in 1997 are expected to be in Grade 7 in 1998. Cohorts are numbered according to the calendar year that they would have been expected to start school; the students mentioned above are referred to as Cohort 1991. Tests by the few students who repeated a grade were removed from the study, as they no longer fitted into just one cohort.

4.2 Data collection

This section discusses the tests used to collect the data as well as the procedures for selecting the sample and conducting the testing. Possible effects on both the students and teachers due to their involvement in this study are also considered.

4.2.1 The test

The Decimal Comparison Test is the only source of data used in this thesis and it consists of one sheet of paper (see section 4.2.3 for the exception) on which are presented pairs of decimal numbers. The instruction at the top of the test is: *For each pair of decimal numbers, circle the one which is LARGER*. Appendix 1 contains copies of both the original test (DCT1) and the newer version (DCT2). While the unusual nature of this diagnostic test was discussed in the previous chapter, the remainder of this section will address the change of test, and summarise the strengths and limitations of this test.

Change of test DCT1 to DCT2

Table 4.1 provides a basic comparison of the two test versions including the years that each test was used in the ARC study. The reasons for changing the test were discussed in section 2.4.2; an item-by-item analysis of the data collected in 1995 and 1996 using DCT1 (see Stacey and Steinle, 1998) revealed that the students allocated to the various groups were not behaving uniformly in their responses to certain items.

Table 4.1: A comparison of DCT1 and DCT2

Details	DCT1	DCT2
Years in use	mid 1995 to end 1996	start 1997 to mid 1999
Testing Periods	TP1 to TP3	TP4 to TP8
Number of tests	Approx 2000	Approx 9000
Total number of items	25	30*
Item types	A,B,C	1,2,3,4,5,6
Order of items	mixed	grouped

** includes 16 items from DCT1*

In DCT1, the three different item types were presented in a mixed order. Due to the increased number of item types in DCT2, however, it was decided that the different types of items would be grouped within the test. This allowed the completed test paper to be classified more quickly by eye than would have been the case otherwise. It consequently enabled results to be returned very quickly to teachers after testing, which the researchers regarded as ethically important.

This rearrangement of items on the test might be expected to affect the results and, in particular, to reduce the number of students in the unclassified group. This would happen if student made their choice on the current item in light of their choice on the previous item; noting similarities between items of the same type. A pilot study (details in Appendix 2) revealed, however, that this effect was minimal and the decision was made to use the paper test with items grouped by type for all subsequent testing.

Strengths and limitations of the DCT

The DCT does not contain questions that assess a student's ability to perform operations involving decimal numbers nor to solve a contextualised problem. While these are undoubtedly important abilities that teachers need to assess, this thesis is concerned with students' understanding of decimal numeration. The comparison task has proved particularly useful for this purpose, as outlined in Chapter 2, but it does have clear limitations.

The main limitation of the DCT is that it is attempting to measure conceptual understanding with a task for which there are two simple algorithms available (referred to as *expert algorithms*). Hence, either a teacher or another student can assist a student to complete the test as an expert. Simply reminding a student of a procedure, such as "compare from left to right" or "add zeros" may assist them to answer all (or most) items correctly. While this is a clear limitation of this test, the considerable number of students that complete tests that do contain errors is a surprise. Furthermore, of the tests completed by students that do contain errors, many fit predicted error patterns as noted in section 3.4. Hence, the DCT cannot distinguish students with a full conceptual understanding of decimal numbers from those who are just following an algorithm not understood; although this limitation could be levelled at many tests.

As noted in Chapter 3, the test cannot distinguish between every known way of thinking. For some ways of thinking, such as *column overflow thinking* and *zero makes small thinking* this is not possible to do with the DCT; it needs an interview or other tasks. For other ways of thinking, however, it is possible to achieve diagnosis by including additional items. Steinle and Stacey (2001) have further modified this test by including a comparison of 0 with 0.6, as well as allowing the equality option, which has enabled new item types (such as 0.8 and 0.800) to be included and hence allowed the identification of other ways of thinking and variations of basic patterns. There are two reasons why DCT2 was used, even because of these inadequacies. One was that some of the variations in students' thinking became evident during the study, when data became available, and another is that it was decided that students would find it too onerous to complete more than 30 items.

4.2.2 Selection of sample

The approximately 3000 students in this study came from 12 schools in the large and geographically spread city of Melbourne. The schools were not chosen randomly but were selected to represent a mix of geographical and socio-economic areas of the city. Because of the commitment required to be involved in the project over several years, volunteer schools were used. Schools where at least one teacher was keen to be a liaison person were first selected. In some cases this was a primary school, and then other primary schools in the area and the local secondary school(s) were also approached to enable the students to be followed after they left Grade 6. In other cases, the first volunteer was in a secondary school and then the local "feeder" primary school(s) were approached. As there was at least one teacher keen to improve within some of the schools in this sample, the prevalence of the misconceptions determined from this volunteer sample may underestimate the prevalence in the general school population.

The effect of other (more subtle) features of the composition of the sample will be discussed throughout the thesis.

While the sample of students was not chosen randomly, we believe that they are a representative sample of schools. To determine whether this is the case, the results will be compared with the 1997 TIMSS-R Australian average on a similar item, (see Chapter 5). However, some findings of this thesis, such as the movement from one misconception to another, are valid even if the schools do not form a sample containing typical proportions in each misconception, because students of a given type have been identified.

To assist in the process of matching students' tests as they pass from primary to secondary school and to examine how socio-economic factors may impinge on results, each group of schools in the same geographical region was allocated a label A to F. A description of the schools is provided in Table 4.2.

Testing was conducted on students in Grades 4 to 10. The lower end of this range was chosen as decimal numeration is typically introduced to students in Grade 4 in Australian schools, although often only to tenths. The upper end of Grade 10 was chosen as it corresponds to the end of compulsory schooling.

Table 4.2: Details of the schools in the sample

School Group	Description and SEIFA*	Grades tested
SGA	One state secondary school and its two feeder primary schools in a low socio-economic area (SEIFA 826)	4 -10
SGB	One private secondary school and its main feeder primary school in a middle socio-economic area (SEIFA 993)	4 -10
SGC	A large state primary school and the two secondary schools to which their students mainly progress in a middle socio-economic area (SEIFA 1063)	4 - 9
SGD	A large state primary school and the secondary school to which their students mainly progress in a middle socio-economic area (SEIFA 1063)	4 - 9
SGE	One private girls' school with both primary and secondary students in a high socio-economic area (SEIFA 1182)	5 - 10
SGF	One private secondary school in a high-middle socio-economic area (SEIFA 1100)	7 - 10

**SEIFA: Socio-Economic Index For Areas, Australian Bureau of Statistics, Index of Relative Socio-Economic Advantage/Disadvantage*

In a longitudinal study such as this, a decision needs to be made regarding the targeting of students for testing as they move from class to class and from school to school, over the four years of the data collection. One option is to target only the students already tested, which would involve inspection of school class lists and isolation of these particular students for the conduct of the test. The second option is to test entire classes that are expected to contain *some* students already in the project. This second option was chosen as it minimised disruption to the school program and it gave useful feedback to the teacher about all their students.

This option of testing entire classes meant that new students could, potentially, be included into the dataset at any time. This proved to be an advantage when the test was altered, as mentioned earlier, at the start of 1997 (from DCT1 to DCT2). For there to be a large number of DCT2 tests completed, many new students were added into the project during 1997. The procedure for identifying tests completed by a particular student is discussed in the section 4.3.1.

4.2.3 Conduct of test and feedback to schools

On some occasions, a researcher involved in the ARC study visited the schools and then was present in the classrooms as the students completed the test. On other occasions, the school indicated that they would prefer the flexibility of the class teacher conducting the testing at a convenient time within, say a 2 week period, and then posting the tests back to the researchers. Usually the test took only a few minutes to administer. All students present in the class at the time of testing completed the test and the tests of those for whom ethics permission was available (nearly 100% since permission had been granted by school principals with parents and students having the right to opt out) were passed to the researchers. No attempt was made to follow up absent students.

Initially, in 1995, students completed this test on a computer, using a mouse to select their answers. While this provided an automatic classification of the test, its disadvantage was the amount of disruption to the class as two students were removed from the class every 5 minutes or so. This method was replaced by the paper test so that the entire class could be tested at once. Preliminary analysis of the data did not detect any differences, so the results from both forms are combined. In both cases, the feedback to the teacher was given reasonably quickly; in some cases, before the researcher left the school.

Information on the diagnosis of each student was given to the class teachers along with brief description of the misconceptions. Teachers could use this information and we will then see the effect of this. Some teachers became more interested and discussed lessons with researchers. The effect on teachers due to their involvement in this project is discussed in section 4.2.5.

4.2.4 Testing schedule

It was the intention of the researchers conducting the ARC study to test students twice per year, from the middle of 1995 to the middle of 1999. This four-year period was divided into eight *testing periods*, as illustrated in Table 4.3. Not all schools tested in all testing periods. For example, School Groups E and F joined the ARC study in Testing Period 3, and School Group A was involved from Testing Period 1 to Testing Period 7, with the exception of Testing Period 3.

Table 4.3: Testing schedule by school group

Testing Period		School Groups						Test Version	Approx number of tests
		SGA	SGB	SGC	SGD	SGE	SGF		
TP1	1995 Semester 2	√	√	√	√			DCT1	2000
TP2	1996 Semester 1	√			√				
TP3	1996 Semester 2		√	√		√	√		
TP4	1997 Semester 1	√		√	√	√		DCT2	9000
TP5	1997 Semester 2	√	√	√	√	√	√		
TP6	1998 Semester 1	√	√	√	√	√	√		
TP7	1998 Semester 2	√	√	√	√	√	√		
TP8	1999 Semester 1		√	√	√	√	√		

There are various reasons why the testing did not always occur at six monthly intervals, or in every testing period, due to the volunteer nature of the study. On one occasion, the liaison teacher at a school delayed the testing until a “more convenient time” but unfortunately this meant that it did not occur within the first half of the year. As a result, there was only one round of testing at that school that year, rather than the two intended. Another complication of the data is that students may be absent on the day of testing, resulting in a gap of 12 months or so between their tests. The amount of time between tests for an individual student will be taken into account in some analyses.

4.2.5 The project effect

Although other elements of the ARC study involved intervention in classrooms (see, for example, Stacey, Helme, Archer & Condon, 2001; McIntosh, Stacey, Tromp & Lightfoot, 2000; Tromp, 1999) there was no project intervention in the schools that provided the longitudinal data. In any longitudinal study, however, the participants are likely to be effected by the repeat testing. In this study, the participants are the students and the teachers. For students, any feelings of uncertainty as they complete the test may result in them seeking additional help or paying closer attention in a future lesson. Thus, participating in the study may help them to realize that there is something that they do not understand. This may result in students improving their knowledge by their next test.

The other participants, the teachers, may have various reactions as well. Some would have ignored the test and its findings, although happy to have their students involved. For others, the fact that a researcher has arrived in the classroom with a test for the students would indicate to the teacher that the task involved is of some importance. According to the liaison teacher at one school, on one occasion the class teachers were comparing the results of their students as if they were "league tables". In this situation, the motivation for a teacher to move their current class "up the league table" in later testing that year (or with new students in the following year) would be very high. Such a situation would be likely to result in a teacher focussing more on the task of decimal comparison before the next round of testing.

A teacher may then focus on this task in one of two ways. The teacher may attempt to increase the students' conceptual understanding by focussing on place value and the meaning of decimal numbers (i.e. *good teaching*). One teacher in School Group D became interested and asked for additional materials, but as this was at the end of the study, this is expected to have had little impact.

The second possibility is that the teacher presents an algorithm for the students to follow which allows them to complete the task correctly with little or no understanding (i.e. *superficial teaching*). It is predicted that the effect of superficial teaching will be expertise that is not permanent. That is, at some later time, the students are likely to forget the rule completely or misapply a half-remembered rule. The phenomenon of a student completing the test as an expert on one occasion but not on the next is referred to as *regression* and will be investigated in Chapter 8. Note that superficial teaching is not only a response to testing.

So, while the intention by the researchers was not to intervene in the events in a normal classroom it is expected that there will be a project effect. Comparisons will be made between the proportions of students in a particular grade, who are experts in their first test, compared with other students in the same grade who are completing a subsequent test. This will give us a measure of the net effect of involvement in the project that may need to be taken into account in determining other measures.

4.3 Procedures in data analysis

The procedures used to analyse the data in this thesis are discussed in this section. These include the tracking of students, and choices made in the analysis and the databases. The aim is to maximise the amount of useful data that can be gleaned from the longitudinal aspects of the study. The intention is to deal with any awkward features of this data rather than discard valuable information. A dataset of maximal size is needed to track some of the rarer ways of thinking and to cover student absences, changes of school, etc.

4.3.1 Tracking students

Nine-digit identification numbers have been assigned to students according to when they completed their *first test* in the ARC study. The first 2 digits identify the school and are followed by the grade level of the first test (from 04 to 10), then the Testing Period (from 01 to 08) and finally a 3-digit number to identify the student.

For example, student 310501023

- was first tested in school 31 (a primary school within School Group C),
- was first tested in Grade 5,
- was first tested in Testing Period 1, and
- was the 23rd such student in this group.

The names of the students have been retained only to assist in matching tests, as described below, and will be deleted from the database at the completion of the study. Ethics approval had been obtained.

To assist with the matching of students' tests, students were allocated to a cohort according to the year that the student would (most probably) have started school. While this category was initially created to assist in matching students' tests, it became a useful way to describe cohorts of students within the data. Table 4.4 indicates how cohorts of students move through various grades in the different testing periods of the study. For example, a student who was tested in Grade 4 in Testing Period 2 (in 1996) would be allocated to Cohort 1992, the most likely year that this student would have started school. These students may have been tested again in Grade 5 (1997), Grade 6 (1998) and Grade 7 (1999) but were not tested beyond that as the data collection finished. A very small number of students were found to repeat a grade, hence changed cohort; tests by these students were deleted.

Table 4.4: Grade of a student in a testing period by cohort

Testing Period	Year	Cohort						
		1992	1991	1990	1989	1988	1987	1986
TP1	1995	*	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9
TP2, TP3	1996	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10
TP4, TP5	1997	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10	*
TP6, TP7	1998	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10	*	*
TP8	1999	Gr 7	Gr 8	Gr 9	Gr 10	*	*	*
Maximum testing opportunities		7	8	8	8	7	5	3

** not tested*

Inspection of the rows indicates that TP2 and TP3 (1996) were the only occasions that tests were completed by students ranging from Grade 4 to 10. In other testing periods a restricted range of grades were tested, as no new Grade 4 students were tested, and as students were followed through to later grades.

Cohort 1986 is the "oldest" group of students in the study. They were in the second semester of Grade 9 at the start of the study and could have been tested at most three times, as the study did not include students in Grades 11 or 12. This pattern of students joining and leaving the project at different times makes data analysis more complicated than a study where all students start and finish as one. This has been undertaken, however, in order to maximize the amount of longitudinal data.

To match tests completed by the same student, the following procedure was implemented. Firstly, information for each test was entered into a row of an electronic file with these details: school group (A to F), first name, surname, grade, and then either detailed information on each test item or a code for the entire test (see section 3.3 for allocation of codes to tests). Then, for each student, the cohort was determined by consulting Table 4.4. The second step was the rearrangement of the data on a sort based on 4 variables in this order: school group, cohort, surname and first name. The third step was the careful examination of the ordered file to check that the matched rows did look suitable before copying the identification number from an earlier test to the current test. This was then supplemented by additional searches in case the first name used by the student changed, for example, from Elizabeth to Liz. Class lists supplied by the school assisted in the identification process. The fourth step was the allocation of new identification numbers to any student who had not been matched by this process. This last step occurred only in TP1 to TP5. After that point, tests from any student not previously in the study were discarded from this database, in order to keep the data set manageable and to concentrate on students who could contribute to longitudinal data.

4.3.2 Analysis decisions

The small numbers of students (n=7) who completed tests which were coded as L3 was especially disappointing as *reverse thinking* was detected in an interview (see Stacey & Steinle, 1998) as well as by other researchers as discussed in Chapter 3. Furthermore, three new items (Type 6) were created for DCT2 especially to detect this way of thinking. This led, before the longitudinal analysis was conducted, to the few L3 tests being recoded as L4 and the code L3 is not mentioned in any further analyses.

A different situation emerged with the S2 tests, which were kept in the longitudinal analyses. As mentioned earlier, the code S2 was created to identify DCT1 tests which appeared to be "in between" S1 and S3 (with a score of 1 out of 2 for Type 5 items), see Table 3.5. It was hoped that later analyses would determine whether these students were more likely to be S1 or S3. For students with an S2 code, their next test was considered.

The distribution of these tests did not match with either of the distributions for tests following S1 or S3. As a result, for most of the analyses that will be reported in this thesis, the codes S2 and S4 have been grouped into S5, although separate results for S2 and S4 exist.

Hence, the codes used in this thesis are those listed in Table 3.3 (with the exception of L3): A1, A2, A3, L1, L2, L4, S1, S3, S5, U1 and U2.

During the analysis of the data, considerable variations were found between the school groups; for example, in the percentage of students answering the tests as task experts (A1). The decision was made to use the full sample of students for most analyses, with variations between the school groups being highlighted where appropriate. In this way, the limitations arising from the sample not being a random sample (e.g. for discussing overall prevalence of codes) is partially overcome by having instead an indication of the variation between schools in different socio-economic areas.

The most desirable level of detail to use in an analysis is semester, as no student can contribute twice to the count, whereas there are many students who will contribute two tests in one grade. Due to the large number of semesters (Semester 1 of Grade 4 to Semester 2 of Grade 10) this level of detail is often contained only in the appendices, with summary figures by school level (i.e. primary school and secondary school) being used within the main chapters to give an overall picture.

A choice that needs to be made in answering the research questions is whether the particular analysis to be conducted should focus on *students* or on *tests*. In a study where a student completes only one test there is no choice to be made. In a longitudinal study, however, issues can often be analysed from both points of view. An example of this is the prevalence of the various codes, dealt with fully in Chapter 5. In this analysis, a focus on *tests* reveals that, for example, 60% of the tests were allocated the code A, while a focus on *students* reveals that 74% of the students were allocated the code A at some time in the study. In most cases, both the test-focussed analysis and the student-focussed analysis have been conducted although only one may be presented in detail.

Initially the data for a test included the class group as well as the school group. This was useful for sorting the many rows within the file and also enabled some observations to be made of the effect of teaching. The class data, however, has not been used in this longitudinal analysis and it is beyond the scope of this study; the wide variation between classes has been documented in Steinle and Stacey (1998a). From this study, it was evident that particular misconceptions sometimes cluster in classes, presumably reflecting either erroneous instruction or misconceptions arising from interference of one part of the curriculum with another as noted in section 2.2.

4.3.3 Database decisions

The statistical package Statview® was initially chosen to record the test data as it has easy-to-use analysis templates. While Statview is effective for cross-sectional analyses, it has proven unsuitable for the longitudinal study due to its inability to treat the information on a student-level rather than test-level.

The statistical package Stata® was designed for longitudinal analyses and has been chosen as a more suitable tool for keeping track of students on repeat testing. After using Stata to identify all the tests belonging to individual students, this information could be stored in various arrays. A preliminary analysis (Stacey & Steinle, 1999a) revealed that the structure of the longitudinal database needs to be carefully considered to allow for students who have missed some rounds of testing but have rejoined the study at a later time.

4.4 Meta-data

Before the main questions of this thesis are investigated the actual testing regime for the students involved in the data collection needs to be established. It was intended that students would be tested twice each year in the ARC Study. The time between tests, as well as the grades that the students were in at the time of testing, and the number of tests done by each student, are all important to the validity of later analyses.

This section provides a useful set of tables that will be referred to throughout the later chapters and illustrates the complexity of the data. For example, every student belongs to one of six school groups and one of seven cohorts. For a particular student, these features do not change, while for any particular test, the grade of the student at the time and the numbers of previous tests already completed by that student, both vary. In the remainder of this chapter the tables and discussion are divided into firstly, a focus on the students and secondly, a focus on the features of the tests.

4.4.1 Statistics focussed on students

All 3204 students involved in the study are discussed in this section. The number of tests completed by each student is described as well as the number of students within each cohort and each school group.

First and last tests for a student

Over one thousand students joined the project in primary school, (i.e. they had their first test in Grades 4 to 6), while another two thousand students joined the project in secondary school (Grades 7 to 10). As well as the first test, Table 4.5 also indicates the last test by every student; of the 1079 students first tested in primary school, 634 (59%) were followed through to secondary school.

Table 4.5: Number of students with first and last tests in each school level.

First test	Last test		Total
	Primary	Secondary	
Primary	445	634	1079
Secondary	0	2125	2125
Total	445	2759	3204

Table 4.6 now provides additional details of the semester when every student completed their first and last test. Again, each of the 3204 students contributes to just one cell in this table; in particular, the 772 students on the main diagonal completed exactly one test. Consider the second row of this table; from the last column we see that 297 students were first tested in Semester 2 of Grade 4. Of these 297 students, 33 completed no further tests (i.e. their first test was their last test) while another 6 (3+3) students had their last tests in Grade 5; and another 92 (11+81) students had their last test in Grade 6. The remaining 166 were followed through to Grades 7 and 8.

Note that it is not possible to determine from this table, the exact number of tests by any student (with the exception of those that lie on the main diagonal and therefore have exactly one test).

Table 4.6: Number of students with first and last tests in each semester

	Semester of first test	Semester of last test														Total
		Primary						Secondary								
		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10		
Primary	Gr4-Sem1	18	0	1	0	3	17									39
	Gr4-Sem2		33	3	3	11	81	70	16	80						297
	Gr5-Sem1			33	10	10	69	112	27	8						269
	Gr5-Sem2				24	1	54	19	11	65	9	27				210
	Gr6-Sem1					28	13	1	11	33	32	31				149
	Gr6-Sem2						33	3	4	12	11	52				115
Secondary	Gr7-Sem1							24	13	10	57	275	37			416
	Gr7-Sem2								94	29	114	216	73	153	52	731
	Gr8-Sem1									10	0	1	9	9	40	69
	Gr8-Sem2										81	19	72	84	60	316
	Gr9-Sem1											5	0	0	17	22
	Gr9-Sem2												106	10	172	288
	Gr10-Sem1													38	0	38
	Gr10-Sem2														245	245
	Total	18	33	37	37	53	267	229	176	247	304	626	297	294	586	3204

In Chapter 5, the number of students who completed tests with a given code while they were in primary school or while they were in secondary school is reported. To assist in this calculation, two separate samples are required; the primary sample consists of the 1079 students who were first tested in primary school, and the secondary sample consists of the 2125 students who were first tested in secondary school. Further details of these samples (for example, the grades of the first and last tests) are provided in Appendix 3.

Total number of tests for a student

As indicated above, there are substantial variations in the length of time that students have been in this study. Table 4.7 indicates the total number of tests that each student has contributed to the study. There are 49 students (2%) who have been tested on 7 occasions while 772 students (24%) have been tested only once. The cumulative frequency column in Table 4.7 shows, for example, that 2432 students (76%) have been tested at least twice and 679 students (21%) have been tested on 5 or more occasions.

While it was intended that students be tested approximately every 6 months there are various logistical reasons why this did not occur. The number of days between the first and last test, summed over all of the 2432 eligible students (i.e. students with at least 2 tests), was 1691605 days. A student with a total of three tests has a *test history* of

[T_1,T_2,T_3], and contributes twice to the count of inter-test periods. The total number of days was divided by the total of the inter-test periods (6658), which gives an overall average inter-test time of 254 days or 8.3 months.

Table 4.7: Distribution of the total number of tests by a student

Number of tests	Number of students		Cumulative Frequency	
7	49	*(2%)*	49	*(2%)*
6	154	*(5%)*	203	*(6%)*
5	476	*(15%)*	679	*(21%)*
4	672	*(21%)*	1351	*(42%)*
3	593	*(19%)*	1944	*(61%)*
2	488	*(15%)*	2432	*(76%)*
1	772	*(24%)*	3204	*(100%)*
Overall	3204	*(100%)*		

Cohorts and school groups

Every student belongs to one of six school groups and to one of seven cohorts as indicated by Table 4.8, although not every cohort was tested in every school group. Different school groups had different patterns of testing; for example, SGA and SGB tested students from all seven cohorts, while SGC and SGD tested only three cohorts.

Table 4.8: Number of students in each school group by cohort

Cohort	School Group						Total	
	SGA	SGB	SGC	SGD	SGE	SGF		
1992	72	33	107	113	43	0	368	*11%*
1991	82	45	117	129	53	0	426	*13%*
1990	147	147	325	155	76	146	996	*31%*
1989	146	140	0	0	86	149	521	*16%*
1988	160	138	0	0	0	146	444	*14%*
1987	64	130	0	0	0	148	342	*11%*
1986	38	46	0	0	0	23	107	*3%*
Total	709	679	549	397	258	612	3204	*100%*
	22%	*21%*	*17%*	*12%*	*8%*	*19%*	*100%*	

The 3204 students are not equally divided between the various school groups. The last row of Table 4.8 indicates that SGE has only 258 students (8% of the sample); SGD has 397 students (12%) while the remaining four have 17% to 22% each. Similarly, the students are not equally divided between the cohorts. The last column of Table 4.8 indicates that Cohort 1990 is the largest cohort with nearly one thousand students (31%) and Cohort 1986 is the smallest with only 107 students (3%).

4.4.2 Statistics focussed on tests

Over eleven thousand tests were completed from 1995 to 1999 and of these, 9862 were identified as belonging to students who joined the study during 1995, 1996 or 1997. (After 1997, a decision was made not to admit any new students to the study, although they may still be tested as classmates of students in the study. These students generated about one thousand tests which were marked and included in the feedback to teachers, but were not entered into the database.) The following section provides details of these 9862 tests, including *when* the tests were completed (both according to the grade of the student at the time and the testing period).

Tests completed in each Grade and Semester

Table 4.9 provides the number and percentage of tests done in each semester and each grade. It can be seen that the number of tests in any given semester is typically between 500 and 1000; the exceptions being the two semesters in Grade 4 with less than this, (39 and 297) and Semester 2 of Grades 7 and 8 with more, (1386 and 1189). The high numbers of tests in Grades 7 and 8 are due to the attempt to follow students from primary school to secondary school where they become classmates of students from other primary schools.

Within the tests in any given semester, there are two factors which affect estimates of the prevalence of the codes; the school group that the students come from and the number of times that the students have already been tested. These will both be discussed below.

Table 4.9: Number and percentage of tests in each grade and semester

	Semester	Number (%) of tests by semester		Number (%) of tests by grade	
Primary	Gr4-Sem1	39	*0.4%*	336	*3%*
	Gr4-Sem2	297	*3%*		
	Gr5-Sem1	414	*4%*	961	*10%*
	Gr5-Sem2	547	*6%*		
	Gr6-Sem1	669	*7%*	1449	*15%*
	Gr6-Sem2	780	*8%*		
Secondary	Gr7-Sem1	911	*9%*	2297	*23%*
	Gr7-Sem2	1386	*14%*		
	Gr8-Sem1	917	*9%*	2106	*21%*
	Gr8-Sem2	1189	*12%*		
	Gr9-Sem1	861	*9%*	1647	*17%*
	Gr9-Sem2	786	*8%*		
	Gr10-Sem1	480	*5%*	1066	*11%*
	Gr10-Sem2	586	*6%*		
	Total	9862	*100%*	9862	*100%*

Firstly, the percentage distribution of tests from each school group is presented in Table 4.10 (Table 3 in Appendix 3 contains the number of tests). The last row indicates that the school group involved the most in testing was SGC (23% of all tests) and SGE was involved the least (11%). The four other school groups each contributed about 15% of tests. Consideration of the total for primary schools reveals that SGC contributed about 1 in 3 tests, and no testing was completed by SGF. It can be seen that SGA was the only school group involved in testing in Grade 4 Semester 1; this testing took place early in the data collection. Furthermore, in Chapter 5 it will be shown that the prevalence of A1 (task expert) was highest in SGE. While SGE contributed to 11% of tests overall, it is over-represented in the testing which took place in some semesters and under-represented in others. This variation in the composition of the tests by school group within a given semester will be important when interpreting the results in later analyses.

Table 4.10: Distribution (%) of tests in each semester by school group

Semester		School Group					
		SGA	SGB	SGC	SGD	SGE	SGF
Primary	**(n=2746)**	**16**	**9**	**36**	**28**	**12**	**0**
Gr4-Sem1	(n=39)	100	0	0	0	0	0
Gr4-Sem2	(n=297)	0	10	60	30	0	0
Gr5-Sem1	(n=414)	30	0	22	38	9	0
Gr5-Sem2	(n=547)	6	12	48	19	15	0
Gr6-Sem1	(n=669)	21	4	28	34	13	0
Gr6-Sem2	(n=780)	11	14	34	24	16	0
Secondary	**(n=7116)**	**17**	**18**	**18**	**12**	**11**	**24**
Gr7-Sem1	(n=911)	14	3	38	31	14	0
Gr7-Sem2	(n=1386)	15	24	19	14	13	15
Gr8-Sem1	(n=917)	14	4	29	21	18	14
Gr8-Sem2	(n=1189)	18	25	16	9	5	27
Gr9-Sem1	(n=861)	12	11	23	13	14	27
Gr9-Sem2	(n=786)	24	31	0	0	7	38
Gr10-Sem1	(n=480)	28	19	0	0	10	43
Gr10-Sem2	(n=586)	25	27	0	0	0	48
Overall	**(n=9862)**	**17**	**15**	**23**	**17**	**11**	**17**

The second factor that may affect estimates of the prevalence of the codes is how often the students have already been tested; see Table 4.11. The last row indicates that 32% of the tests were completed by students with no previous tests (which corresponds to every student's first test), while 25% of the tests were completed by students with exactly one previous test (which corresponds to students' second tests).

For the 297 tests completed in Semester 2 of Grade 4, for example, all students (100%) were on their first test. Contrast this with the 917 tests completed in Semester 1 of Grade 8. Only 8% of these tests were completed by students with no previous tests, while 35% had exactly one previous test, and 29% had exactly 2 previous tests. In fact, 4% of these tests were completed by students with 6 previous tests, i.e. these students were completing their 7^{th} tests. Hence, there is considerable variation, between the semesters, in the numbers of previous tests completed by the students. While overall, students who had completed no earlier test did 32% of tests, this varies considerably from a low of 3% (Semester 1 of Grade 9) to the 100% mentioned above (Semester 2 of Grade 4).

Table 4.11: Distribution (%) of the number of previous tests in each semester

Semester (number of tests)	Number of previous tests							Average
	0	1	2	3	4	5	6	
Gr4-Sem1 (n=39)	100							0.0
Gr4-Sem2 (n=297)	100							0.0
Gr5-Sem1 (n=414)	65	35						0.4
Gr5-Sem2 (n=547)	38	47	14					0.8
Gr6-Sem1 (n=669)	22	25	42	11				1.4
Gr6-Sem2 (n=780)	15	20	23	34	8			2.0
Gr7-Sem1 (n=911)	46	10	9	14	18	4		1.6
Gr7-Sem2 (n=1386)	53	25	8	5	6	3		1.0
Gr8-Sem1 (n=917)	8	35	29	10	7	7	4	2.1
Gr8-Sem2 (n=1189)	27	29	21	18	4	2		1.5
Gr9-Sem1 (n=861)	3	15	29	25	22	5	1	2.7
Gr9-Sem2 (n=786)	37	29	20	14				1.1
Gr10-Sem1 (n=480)	8	26	32	20	13			2.0
Gr10-Sem2 (n=586)	42	20	23	14	1			1.1
Overall (n=9862)	32	25	20	14	7	2	0.5	1.5

As well as the distribution, the average number of previous tests has been calculated and provided in the last column of Table 4.11. There are several entries of 2 or more, for example, Semester 2 of Grade 6, and Semester 1 in Grades 8, 9 and 10. This may well be accompanied by higher than normal levels of expertise. So, throughout most of the testing in the secondary schools, the students tested in the second half of the year have been tested less than the students tested in the first half. If a (positive) project effect exists, then students in the first half of the year may well appear to be "cleverer" than expected.

As a final note on the variation between the testing in the various school groups, Table 4 in Appendix 3 contains the average number of previous tests in each semester for the different school groups (i.e. the last column of Table 4.11 has been calculated for each school group separately). While the average number of previous tests for students tested in Semester 1 of Grade 10 is 2.0 (for the full sample of 480 students), there are considerable differences between the school groups. School Group A has an average of 1.2 (n=133), while School Group E has an average of 3.6 (n=50). This highlights again the need to carefully consider any project effect and variations between school groups in any analysis.

Tests completed in each Testing Period

Table 4.4 indicated how the movement of students through grades over time is more easily described by referring to cohorts of students. In Table 4.12, the number of tests done in each of the eight testing periods by students from each of the seven cohorts is indicated, as well as the grade of the student at the time of testing (shaded entries).

Table 4.12: Number of tests by testing period and cohort

Testing Period	Cohort														Total
	1992		1991		1990		1989		1988		1987		1986		
TP1			173	*4*	83	*5*	0	*6*	122	*7*	24	*8*	30	*9*	432
TP2	39	*4*	127	*5*	102	*6*	82	*7*	76	*8*	22	*9*	38	*10*	486
TP3	124		171		167		232		100		95		42		931
TP4	287	*5*	286	*6*	440	*7*	73	*8*	0	*9*	0	*10*			1086
TP5	293		318		814		383		350		283				2441
TP6*	281	*6*	206	*7*	581	*8*	259	*9*	217	*10*					1544
TP7*	295		218		682		311		261						1767
TP8*	183	*7*	187	*8*	580	*9*	225	*10*							1175
Total	1502		1686		3449		1565		1126		424		110		9862

Numbers in shaded cells indicate the grade at the time of testing
** no new students admitted in TP6 to TP8, so these tests are a subset of the number actually done*

The last row reveals that Cohort 1990 provided the most tests (over 3000) and it is also the largest cohort with nearly one thousand students (see Table 4.8). Cohort 1986 has the least number of tests (110 tests) and is the smallest cohort (107 students). Nearly two and a half thousand tests were completed in TP5; in fact, additional tests were completed in TP6 to TP8 but are not included as no new students were added to the dataset after TP5. From TP6 to TP8 the focus was on building the longitudinal data.

Table 4.13 contains similar information as Table 4.12 (the last columns are identical) but in Table 4.13 the columns indicate the school group involved in the testing, rather than the cohort. Note that SGE and SGF joined the study in TP3 and that not every school group tested in every testing period.

Table 4.13: Numbers of tests by testing period and school group

Testing Period	School Group						Total
	SGA	SGB	SGC	SGD	SGE	SGF	
TP1	48	128	166	90	-	-	432
TP2	383	0	0	103	-	-	486
TP3	0	309	272	0	159	191	931
TP4	102	0	440	327	217	0	1086
TP5	508	494	409	319	148	563	2441
TP6	314	27	349	287	203	364	1544
TP7	316	302	324	272	194	359	1767
TP8	0	243	297	256	172	207	1175
Total	1671	1503	2257	1654	1093	1684	9862

- indicates school group had not yet joined study

The students from SGC provided over two thousand tests, while the least number of tests came from SGE (just over one thousand tests). Typically more than two hundred tests were completed within any school group in a testing period.

For completeness, an analogous table to Table 4.8 is now included. Table 4.14 has the same rows and columns as Table 4.8, but has the number of tests rather than number of students.

Table 4.14: Number of tests in each school group by cohort

Cohort	School Group						Total
	SGA	SGB	SGC	SGD	SGE	SGF	
1992	209	132	512	459	190	0	1502
1991	242	148	486	560	250	0	1686
1990	321	387	1259	635	321	526	3449
1989	352	377	0	0	332	504	1565
1988	428	254	0	0	0	444	1126
1987	81	156	0	0	0	187	424
1986	38	49	0	0	0	23	110
Total	1671	1503	2257	1654	1093	1684	9862

Appendix 3 contains a series of six tables (Tables 5 to 10) which provide, for each school group, the number of tests in each semester by each cohort.

4.4.3 Statistics focussed on both students and tests

Table 4.15 contains the distribution of tests by students within each school group and the average number of tests per student. While the average number of tests per student is 3.1 (see last row), none of the school groups can be thought of as “average”. Rather, there are three school groups that have between 2 and 3 tests per student (i.e. SGA, SGB and SGF) and the remaining three (SGC, SGD and SGE) have an average of more than 4 tests per student.

Table 4.15: Distribution of the total number of tests for students by school group

School Group	Total number of tests for a student							Total students	Total tests	Average
	1	2	3	4	5	6	7			
SGA	232	167	167	111	32			709	1671	2.4
SGB	255	148	171	86	19			679	1503	2.2
SGC	53	55	53	131	160	68	29	549	2257	4.1
SGD	40	29	45	83	131	49	20	397	1654	4.2
SGE	24	12	21	60	104	37		258	1093	4.2
SGF	168	77	136	201	30			612	1684	2.8
Total	772	488	593	672	476	154	49	3204	9862	3.1

Table 4.16 contains similar information (the last rows are identical) but this time the rows are cohorts rather than school groups. The lowest figures are from Cohorts 1987 and 1986, which had fewer testing opportunities (see Table 4.4).

Table 4.16: Distribution of the total number of tests for students by cohort

Cohort	Total number of tests for a student							Total students	Total tests	Average
	1	2	3	4	5	6	7			
1992	37	16	39	97	146	33		368	1502	4.1
1991	62	43	63	87	62	72	37	426	1686	4.0
1990	120	157	194	265	199	49	12	996	3449	3.5
1989	85	91	144	139	62			521	1565	3.0
1988	102	100	151	84	7			444	1126	2.5
1987	262	78	2					342	424	1.2
1986	104	3						107	110	1.0
Total	772	488	593	672	476	154	49	3204	9862	3.1

Table 4.8 contains the number of students in each school group and cohort, while the number of tests for these students is provided in Table 4.14. The ratios of these quantities are provided in Table 4.17.

Table 4.17: Average number of tests per student in each school group by cohort

Cohort	School Group						Average
	SGA	SGB	SGC	SGD	SGE	SGF	
1992	2.9	4.0	4.8	4.1	4.4		4.1
1991	3.0	3.3	4.2	4.3	4.7		4.0
1990	2.2	2.6	3.9	4.1	4.2	3.6	3.5
1989	2.4	2.7			3.9	3.4	3.0
1988	2.7	1.8				3.0	2.5
1987	1.3	1.2				1.3	1.2
1986	1.0	1.1				1.0	1.0
Average	2.4	2.2	4.1	4.2	4.2	2.8	3.1

Note the different patterns of testing between the school groups. For example, SGA and SGB have students from all cohorts while SGC and SGD have fewer students in total, but more from the three cohorts that they do test. So, while testing in SGC and SGD was concentrated on students in just a few cohorts, testing in SGA and SGB was more spread across various cohorts, which results in the fewer tests per student, found earlier.

The implications of these sampling fluctuations are discussed throughout subsequent chapters.

4.5 Conclusion

The ARC study collected a large amount of data over a period of four years. As a volunteer sample was used, albeit a very large one, care will be needed in making any generalisations for the school population at large. The effect, on both the students and the teachers, of being involved in the project has been discussed and will also need to be considered when results are being interpreted. The testing schedule was not as consistent as would be desirable, but this was due to the volunteer nature of the sample;

a key person at each school was the liaison teacher who was relied on to encourage the teachers at their school.

At the same time, however, it is clear that a large and rich dataset has been obtained. Over three thousand students from 12 schools (representing a wide socio-economic mix) were involved and they completed nearly ten thousand tests, with an average inter-test time of 8.3 months. Nearly 60% of the one thousand students who were first tested in primary school (Grades 4 to 6) were tracked to their secondary school. More than 600 students completed 5 or more tests. The complications in the dataset due to students joining and leaving the study at different times and grades have been documented, as have the different testing schedules for the different school groups. These will be taken into consideration in the following chapters when the data is being analysed.

CHAPTER 5 GENERAL DATA ANALYSIS

5.1 Introduction and terminology

The analysis in this chapter contains a relatively straightforward interrogation of the data with respect to two features; the *prevalence* of the various codes and the *transitions* between the codes on consecutive tests. Also included is an introduction to *students' test histories*. (These terms are defined carefully below, as well as in the Glossary).

For some issues, a more sophisticated analysis is warranted, which is better able to deal with irregularities arising from sampling, potential effects of being involved in the project, etc. This chapter therefore provides a first description and sets the scene for the analyses that are presented in Chapters 6 to 9.

Terminology

Various definitions for three types of *test-focussed prevalence* (overall, school level and semester) and for two types of *student-focussed prevalence* (overall and school level) are provided in the next section.

The *project effect* is defined to be the improved performance on the DCT which is due to having completed the test before. It is calculated by determining the prevalence of A1 (*task expert*) for two groups of students in the same semester (those on their first test and those on a *subsequent* test) and then subtracting these two figures.

Each student has an (ordered) list of codes indicating the tests completed by that student: this is referred to as that student's *test history*. In general a student who has completed a total of three tests has a test history of [T_1,T_2,T_3] but a specific example might be [L,U,A] using the coarse codes or [L1,U1,A2] using the fine codes.

Pairs of *consecutive* tests within student's test histories are examined to provide information about the *transitions* or movements between codes. For example, a student who has completed a total of three tests will contribute two pairs of consecutive tests: (T_1, T_2) and (T_2, T_3).

Section 5.2 contains the prevalence (from both test-focussed and student-focussed perspectives) of the various codes and discusses variations with age and with school group. The existence of a *project effect* will be investigated; it is expected that the repeated testing of students and the provision of results to the class teachers increase the prevalence of expertise. This project effect may need to be considered in later chapters and will determine which tests are considered in determining improved estimates of the prevalence (both test-focussed and student-focussed perspectives).

Section 5.3 introduces students' test histories; tracing the changes in the codes from one test to another should reveal changes in that student's thinking. Section 5.4 considers the transitions students make by analysing pairs of consecutive tests. This is done using both the coarse codes and the fine codes. Table 5.18 is a foundation table of this chapter and is the basis for later investigations in Chapters 6, 7 and 8.

5.2 Prevalence of codes

In this section the prevalence of the coarse and fine codes will be presented. Note that Steinle and Stacey (1998a) used the term *incidence*, but in keeping with the medical literature, we have changed to *prevalence* in subsequent publications. The following definitions from the Oxford Handbook of Clinical Medicine (Hope, Longmore, McManus & Wood-Allum, 1998) have been adapted for this thesis. "The *period prevalence* of a disease is the number of cases, at any time during the study period, divided by the population at risk. … *Point prevalence* is the prevalence at a point in time", (p741). In other words, the *period prevalence* of a disease is the total number of cases reported during the study period, divided by the population of interest. The *point prevalence* is the total number of cases at a particular point of time, divided by the population of interest. The following example illustrates the difference between the period and point prevalence. Imagine a survey of 1000 people that included these four questions:

- Do you have the hiccups at this very moment?
- Have you had the hiccups in the last week?
- Have you had the hiccups in the last year?
- Have you ever had the hiccups?

If, as is likely, almost everyone answered the first question "No", then the prevalence of hiccups at that particular point in time (point prevalence) is almost 0%. Some people may answer in the affirmative to the second question, however, giving a weekly prevalence of, say, 3%. The yearly prevalence will be at least 3% as additional people join those who answered in the affirmative to the second question. Hence, the yearly prevalence may be 20%, but cannot be predicted from the weekly prevalence.

The response to the final question could easily be 100%; the *lifetime prevalence* or *lifetime risk* is then 100%. So, hiccups affect almost everybody at some point in time, (i.e. the lifetime prevalence is very high) even though the point prevalence may be very small. Note that none of these results can be determined from the others, even if exemplary sampling is used. The only relationship is that as the period increases, the number of people in the sample who indicate that they were involved in the particular phenomenon either remains the same or increases, and hence there is a non-decreasing sequence of results:

point prevalence ≤ *weekly prevalence* ≤ *yearly prevalence* ≤ *lifetime prevalence*

A cross-sectional study can only provide the point prevalence, while following a student over a period of time in a longitudinal study provides an indication of a period prevalence. As discussed below, for the present study, the most useful measure of point prevalence is the percentage of tests allocated to each code in a given semester (e.g. Semester 1 of Grade 7), and the most useful measures of period prevalence are the percentage of students allocated a code over the period of time that they are in primary or secondary school. Note that the more observations taken on the student, within the given period, the more accurate the result for the period prevalence. For example, if a fixed group of students were tested repeatedly over a fixed period of three years (say Grades 4, 5 and 6), the period prevalence based on more tests is *at the least the result* from fewer tests:

prevalence (two tests) ≤ *prevalence (three tests)* ≤ *prevalence (four tests)*

These considerations lead to the following definitions, where the concepts above are adapted for the testing regime of the longitudinal data. Some of these quantities are affected more than others by the composition of the sample and the testing regime within each school group. As will be shown later it is useful to classify these quantities according to whether they focus on tests or on students.

Definitions for Test-Focussed Prevalence (TFP)

The *overall TFP* of a code is calculated by dividing the number of tests allocated this code in this study by the total number of tests completed in this study (expressed as a percent). This is a useful single summary quantity but as it combines tests from students in Grade 4 to Grade 10, it smooths over differences between younger and older students. Note that almost every student contributes several tests to this quantity.

The *school level TFP* of a code is calculated by dividing the number of tests allocated this code in a given school level (Primary or Secondary) by the total number of tests completed in the same school level (expressed as a percent). These two quantities, Primary TFP and Secondary TFP are reasonable indicators for teachers as they are based on students from Grades 4 to 6, and Grades 7 to 10, respectively. They are also useful quantities in this thesis as a compromise between too much and too little detail and they avoid small sample sizes.

The *semester TFP* of a code is calculated by dividing the number of tests allocated this code in a given *semester* (for example Semester 2 of Grade 7) by the total number of tests completed in the same *semester* (expressed as a percent). This is the quantity that corresponds most closely to the point prevalence, because testing was carried out at most once per semester. Due to the large number of semesters involved in this study, (Semester 1 of Grade 4 to Semester 2 of Grade 10), as well as the number of codes to be investigated, details of this measure will normally only be provided in appendices and the school level TFP (defined above) will be used as a useful summary.

Hence, the semester TFP provides the prevalence of the various codes at a point in time (point prevalence); the period prevalence (obtained by identifying students and following them over a period of time) will now be discussed.

Definitions for Student-Focussed Prevalence (SFP)

The smallest period over which students will be followed is the school level (either Primary or Secondary school) as several tests are required to observe their thinking and its changes. For this particular measure, students are allocated to either the Primary sample or the Secondary sample depending on their first tests and, by definition, are not followed from Primary to Secondary.

The *school level SFP* of a code is calculated by dividing the number of students who were allocated a given code, (on any test that they completed at any time while they were at this particular school level – Primary or Secondary), by the number of students in this particular school level (i.e. the population at risk), expressed as a percent. As above, these two quantities, Primary SFP and Secondary SFP have appropriate data sets for the data analysis and have the added advantage that they are reasonable indicators for teachers as they are based on the familiar patterns of school organisation. As mentioned above, students who are tested less have fewer opportunities to reveal how they are thinking and so these students dilute these measures. Hence, the school level SFP calculated from this dataset is an underestimate of the percentage of students who are involved in each code while in primary or secondary school.

The *overall SFP* of a code is calculated by dividing the number of students who were allocated this code, (on any test that they completed at any time during the course of the study), by the number of students in the study (expressed as a percent). This summative quantity is useful for initial investigations but as no students were followed from Grade 4 to Grade 10, this quantity does not represent the "lifetime risk" of a student being involved in a certain code. If there was a fixed sample of students who were followed from Grade 4 to Grade 10 then this quantity would be greater than both the measures created for the shorter time periods (school levels) but, as will be seen, this is not the case here.

In summary, the TFP is affected by the representativeness of the students present on a given day of testing, while the SFP is affected by their presence over the whole testing regime.

Section 5.2.1 provides the overall TFP of the codes. Further investigation of the TFP for each code, by considering the influence of school level and school group is then reported. Section 5.2.3 contains similar details for the student-focussed prevalence (SFP) for each code.

5.2.1 Test-focussed prevalence (TFP)

Table 5.1 lists the overall Test-Focussed Prevalence (TFP) of both the coarse and fine codes when all 9862 tests are considered. The most frequently occurring coarse code is A with 5930 (60%) tests. The next most frequent code is L, which has been allocated to 16% of the tests, and then S (12% of the tests). This leaves another 12% of the tests being assigned the code U as they did not match the criteria for A, L or S (see Tables 3.2 and 3.3 for the allocation of the coarse and fine codes, respectively).

Table 5.1: Overall TFP of both coarse and fine codes

Coarse code	Tests		Fine code	Tests	
	Number	Prevalence		Number	Prevalence
A	5930	60%	A1	5214	53%
			A2	418	4%
			A3	298	3%
L	1596	16%	L1	1066	11%
			L2	298	3%
			L4	232	2%
S	1143	12%	S1	316	3%
			S3	526	5%
			S5	301	3%
U	1193	12%	U1	1105	11%
			U2	88	1%
Total	9862	100%		9862	100%

By far the most frequent fine code is A1 (*task expert*) being allocated to over half of all tests. After A1, the next most frequently occurring fine codes are L1 (*whole number thinking*, 11%) and U1 (unclassified tests, also 11%). The remaining eight fine codes range from 1% to 5% and might not appear, from these figures, to be of much educational importance; however, other ways of looking at the data below show that they affect many students.

The fine code A1 dominates the coarse code A; of the tests coded as A, nearly 90% were A1. Similarly over 90% of the U codes are U1. The coarse code L is also dominated by one fine code (L1, 67%), but not to the same extent. The situation with S is very different; S3 is the most frequently occurring S code (about half of the S tests) but both S1 and S5 make considerable contributions to the overall S code; about one quarter each. This feature is an early indication of the unusual nature of the S codes. The fine code U2 can be seen to be very rare. The discussion below will indicate how well these overall figures represent students of various ages and from various school groups.

TFP by School Level

Before more detailed results are discussed, a general grouping into school level is provided in Table 5.2. Tests completed by students in Grades 4 to 6 are allocated to Primary level, while tests completed by students in Grades 7 to 10 are allocated to Secondary level. An addition column gives the difference between the Secondary TFP and Primary TFP; a positive entry denoting a higher prevalence in secondary schools. The previous overall prevalence of A (60%) can be seen to be a combination of a lower figure in primary school (36%) and a higher figure in secondary school (69%), as would be expected; likewise for A1. Younger students, in contrast, are more likely to be involved in L and L1 than older students. The prevalence of both S and U appear to be more constant.

Table 5.2: School level TFP of both coarse and fine codes

Coarse code	Primary TFP (n=2746)	Secondary TFP (n=7116)	Secondary-Primary	Fine code	Primary TFP (n=2746)	Secondary TFP (n=7116)	Secondary-Primary
A	36	69	+33	A1	31	61	+30
				A2	3	5	+2
				A3	2	3	+1
L	37	8	-29	L1	27	4	-23
				L2	5	2	-3
				L4	5	1	-4
S	13	11	-2	S1	4	3	-2
				S3	6	5	0
				S5	3	3	0
U	14	12	-2	U1	12	10	-3
				U2	0	1	+1
Total	*100*	*100*	*0*		*100*	*100*	*0*

At this point, it should be noted that only 60% of the tests completed in secondary schools were allocated the code A1 (*task expert*). Of the remaining 40%, approximately 10% were answered inconsistently which leaves approximately 30% which showed evidence of consistent errors. This highlights the significance of this research into students' misconceptions.

TFP by School Group

Now consider variations between school groups. Table 5.3 contains the Primary TFP and Secondary TFP by each school group for the coarse codes, and Table 5.4 contains similar information for the fine codes. (Note that the sum of the rows is 100% except for rounding errors.)

Table 5.3: TFP of coarse codes by school level and school group

School Level and School Group		Coarse codes			
		A	L	S	U
Primary	**(n=2746)**	**36**	**37**	**13**	**14**
SGA	(n=429)	12	64	8	16
SGB	(n=235)	34	31	27	9
SGC	(n=987)	30	44	13	14
SGD	(n=768)	46	27	14	13
SGE	(n=327)	67	12	7	14
Secondary	**(n=7116)**	**69**	**8**	**11**	**12**
SGA	(n=1242)	56	16	13	15
SGB	(n=1268)	69	7	13	11
SGC	(n=1270)	68	10	9	12
SGD	(n=886)	72	6	12	10
SGE	(n=766)	87	2	7	5
SGF	(n=1684)	71	5	11	13

Note that all the fine codes are found in all of the school groups; even U2, although sufficiently infrequently to be less than 0.5% in primary schools and hence not evident in the table. Also note that there are some considerable variations between school groups. For example, the primary school in SGB has two to three times more tests being allocated S codes than the other primary schools; over one in four tests from this primary school are allocated the code S.

Table 5.4: TFP of fine codes by school level and school group

School Level and School Group		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
		Fine codes										
Primary	**(n=2746)**	**31**	**3**	**2**	**27**	**5**	**5**	**4**	**6**	**3**	**13**	**0**
SGA	(n=429)	8	2	2	54	5	5	2	3	3	15	0
SGB	(n=235)	29	3	2	24	3	4	7	13	6	9	0
SGC	(n=987)	24	2	3	30	6	7	5	5	3	13	0
SGD	(n=768)	40	3	2	18	5	4	5	6	3	13	0
SGE	(n=327)	64	2	1	6	3	3	2	2	2	14	0
Secondary	**(n=7116)**	**61**	**5**	**3**	**4**	**2**	**1**	**3**	**5**	**3**	**10**	**1**
SGA	(n=1242)	48	5	3	10	4	3	2	6	4	14	1
SGB	(n=1268)	61	4	4	4	2	1	4	5	4	9	1
SGC	(n=1270)	61	4	3	6	3	2	2	4	2	11	1
SGD	(n=886)	65	4	3	3	2	1	3	6	2	10	1
SGE	(n=766)	81	4	2	1	1	0	2	4	2	4	1
SGF	(n=1684)	60	7	4	2	2	1	3	5	3	11	1

Using the prevalence of A1 in Table 5.4 as an indicator, the school groups fall into three categories; SGA is the weakest, SGE is the strongest, and the remaining four school groups fall into an intermediate group. In particular, SGA has low prevalence of A and A1 at both the primary and secondary levels, corresponding to high prevalence of both L and L1. SGE presents the opposite story: high prevalence of A and low prevalence of L. Although not a topic within this thesis it should be noted that, within the school groups involved in this study, SGE has the highest socio-economic status and SGA the lowest (see Table 4.2). Hence, general student achievement is the most likely explanation of the difference rather than any factor associated with this particular study, such as the varying testing regime within the different schools.

Another explanation for the variation in the prevalence of A and A1 might be that some school groups contain older or younger students; this will now be considered.

TFP by semester

Consideration of the broader picture (school level) is smoothing any clustering of the codes, which might be due to teaching; hence more detailed information (by semester) will be summarised below. (The full set of results is provided in Appendix 4.)

Generally, the prevalence of A1 increases with each semester, from less than 10% of students in Grade 4 to about 75% of students in Grade 10. There is an apparent dip in the prevalence of A1 in Semester 2 of each of the Grades 7 to 10, which will be shown later to be due to the project effect. (In other words, students who have completed more tests are better than expected at the task.) As noted above, there is considerable variation in the prevalence of A1 between the school groups and this is clearly shown in Figure 5.1, (which is created from the data in Table 3 in Appendix 4, by considering only the figures based on 50 or more tests).

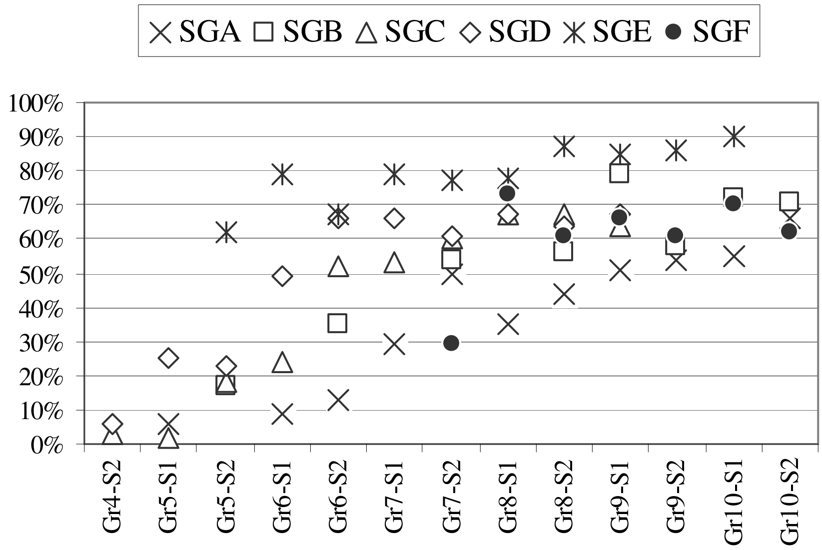


Figure 5.1: Prevalence of A1 by semester for all tests, by school group

In Chapter 4 it was noted that the sample for this study was not randomly selected, but it was claimed to be a representative sample. In order to confirm this claim, the prevalence of A1 for the Grade 8 students in this sample will now be compared with the TIMSS-R (1997) results for the Australian sample. Table 2.3 indicated that 58% of the Grade 8 students in the Australian sample were able to choose correctly on the TIMSS-R task. Figure 5.1 indicates considerable variation between the school groups in both semesters of Grade 8, which certainly span 60%. Further detail (obtained from Table 3 in Appendix 4) indicates the overall results of 65% and 59% for the two semesters in Grade 8, which is consistent with the TIMSS-R result.

Other comparisons with the Australian TIMSS-R sample suggest that the Grade 8 students within the sample for this study have a lower prevalence of both L and S behaviours (see Tables 6 and 10 in Appendix 4). This apparent difference will be resolved in Chapter 9, when the improved test-focussed prevalence of the codes are presented.

Other important results from Appendix 4 are now summarised:

- The prevalence of the code A2 increases slowly from 1% in Grade 4 to 6% in Grade 10, and a similar trend is observed with the code A3. The total of these two codes will be shown later to be of interest; four of the school groups have a total of 10% or more of these codes in particular semesters.
- As might be predicted by the observed drop in the prevalence of L1 from primary to secondary, there is a general decrease in L1 over the semesters. There are considerable differences between the school groups. For example, by the start of Grade 7, about 25% of students in SGA tested as L1 compared with only 2% of students in SGE. Thus the school group with the highest prevalence of A1 has the lowest prevalence of L1 and vice-versa.
- The prevalence of both L2 and L4 decrease similarly with grade.
- The codes S1, S3 and S5 each have a reasonably flat distribution of 5% or less. The S3 code, however, appears to be slightly more prevalent (7%) in both semesters in Grade 5 and both semesters in Grade 8. There are several school groups with much larger figures; 17% of the students in Semester 2 of Grade 7 in SGF have S3 codes, as do 15% of the students in Semester 2 of Grade 5 in SGB. The code S3 will receive more investigation later this thesis.
- The code U1 is slightly more prevalent in Semester 2 of Grade 5, (17%) but otherwise is typically 9% to 14%. The code U2 is quite rare, but appears to be slightly more concentrated in SGF (up to 4% in semester 1 of Grade 10).

Non-A1 codes

Given a student completes a test that is not coded as A1, what is the distribution of these non-A1 tests? Table 5.5 provides this distribution by semester. In primary schools, L1 is clearly the most likely non-A1 code, while in secondary schools it is U1.

As expected, L1 decreases with semester, and A2 and A3 increase. Two other codes have interesting peaks, S1 accounts for 12% of the non-A1 tests in Semester 2 of Grade 7, and S3 accounts for between 17% and 20% of the non-A1 tests in Grade 8. These unexpected peaks in S1 and S3 in junior secondary school are consistent with the claim that these ways of thinking are at least partially explained as a result of the interference from new teaching, as suggested in section 2.2. This is discussed further in Chapter 9.

The figures in Table 5.5 will be used in the investigation of regression in Chapter 8. They provide the background prevalence of the non-expert codes which will be compared with the codes of the tests which indicate regression from A1.

Table 5.5: Distribution (%) of non-A1 tests by semester

	Semester	Fine codes (non-A1)									
		A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	Gr4-Sem1 (n= 39)	0	0	82	3	0	0	0	5	10	0
	Gr4-Sem2 (n= 285)	1	2	65	5	7	2	4	3	12	0
	Gr5-Sem1 (n= 356)	2	3	49	6	7	6	9	4	15	0
	Gr5-Sem2 (n= 412)	4	3	35	8	8	7	9	4	23	0
	Gr6-Sem1 (n= 416)	5	3	32	9	7	8	7	5	23	0
	Gr6-Sem2 (n= 384)	7	7	21	8	8	8	11	6	24	0
Secondary	Gr7-Sem1 (n= 382)	9	8	25	8	4	7	9	5	24	0
	Gr7-Sem2 (n= 624)	11	6	15	6	5	12	13	8	22	1
	Gr8-Sem1 (n= 320)	12	7	8	8	3	5	20	8	26	4
	Gr8-Sem2 (n= 486)	13	8	10	6	3	5	17	9	24	4
	Gr9-Sem1 (n= 275)	13	11	6	3	2	7	12	6	35	4
	Gr9-Sem2 (n= 313)	16	10	5	5	2	7	12	10	29	4
	Gr10-Sem1 (n= 152)	19	13	6	2	3	5	8	9	28	7
	Gr10-Sem2 (n= 204)	13	12	5	3	2	3	11	8	37	5
	Overall (n= 4648)	9	6	23	6	5	7	11	6	24	2

The apparent drop in expertise (A1) in Semester 2 of each grade in secondary school (see Table 3 in Appendix 4) will now be investigated before the student-focussed prevalence (SFP) is discussed.

5.2.2 The project effect

The previous results for the TFP of the codes were based on all tests completed in each semester. This section tests the hypothesis of section 4.2.5 that there will be a positive project effect; i.e. students who have done the test before are better than other students of the same age. Figure 5.2 presents the prevalence of A1 by semester for two groups; namely, students completing their *first test* and students completing a *subsequent test*. So, all 9862 tests contribute to the calculations of the prevalence of A1 in this figure; 3204 tests are involved in the *first test* calculations (every student's first test) and 6658 tests are involved in the *subsequent test* calculations. (Note that any student contributes at most one test in any given semester, and therefore the calculation of the differences between the two graphs and the chi-squared analysis are appropriate.)

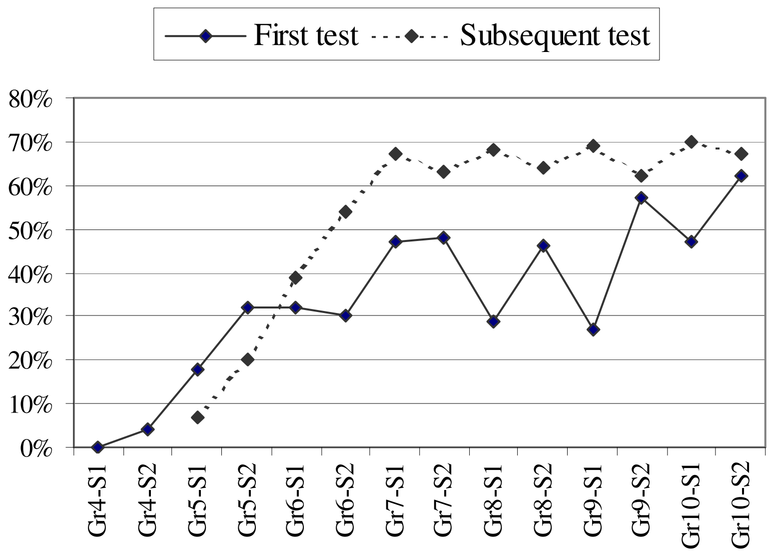


Figure 5.2: Prevalence of A1 by semester for two samples of students (first tests and subsequent tests)

It is clear that for students in Grades 6 to 10, there is a (positive) project effect and for students in Grade 5, there is a (negative) project effect. A chi-squared test (see Appendix 11) proves that the difference between the prevalence of A1 for the two groups is significant ($p<0.05$) for nine of the twelve semesters from Grade 5 to Grade 10. Only three semesters (Semester 1, Grade 6 and Semester 2 in both Grades 9 and 10) have a difference that is not significant at the 0.05 level. The negative project effect in Grade 5 will be explained when regression is investigated in Chapter 8.

This evidence of a project effect supports the explanation that retesting improves performance because of a combination of students redoing the test and teachers placing some emphasis on this skill, as well the fact that the teachers were given feedback after testing (see section 4.2.3).

Hence, overall the test effect has been found to be positive, which explains some of the variations in the levels of expertise; students who have completed more tests are more likely to be experts than other students (at the same grade and semester) who are on their first test. Furthermore, the TFP of the code A1 *(task expert)* found in the previous section are overestimates of the ability of the general population. Likewise, the severity of the problems is underestimated by these earlier results.

Figure 5.2 also shows clearly a regular rise and fall in the prevalence of A1 that is due to combining data from the various school groups. For example, SGE has been noted as a school with high levels of expertise, and it contributes only 5% of the tests by students in Semester 2 of Grade 8, (compared with 18% in Semester 1). It is therefore not a surprise that the (average) levels of expertise in Semester 2 are lower than in Semester 1. Note that this is because it is *not* the same students completing both tests in Grade 8; students from five different cohorts (1987 to 1991) completed tests in Grade 8.

An attempt was made to determine the size of the project effect as follows. For each semester, the average difference between the prevalence of A1 for the two samples was determined (given that there were at least 50 tests involved in both calculations). It was found to be approximately 10%. For the purposes of this thesis, no numerical value for the project effect is required; rather the presence of the project effect highlights the need for very careful selection of samples for later analyses.

This concludes the discussion of the test-focussed prevalence (TFP) of the coarse and fine codes; the next section will consider the student-focussed prevalence (SFP).

5.2.3 Student-focussed prevalence (SFP)

As mentioned earlier, another way of determining the significance of the various codes is to consider the number of students who tested a certain way at some stage of the study. Steinle and Stacey (2002) provided details of the student-focussed prevalence (SFP) of both A2 and U2 (11% and 2%, respectively) using the term *longitudinal incidence.*

The overall student-focussed prevalence (SFP) for each code (for the 3204 students in the study) is given in Table 5.6. Note that, as students often contribute to several rows, the column sum of the percentages exceeds 100%.

Table 5.6: Overall SFP of both coarse and fine codes

Coarse code	Students (n=3204) Number	Prevalence	Fine code	Students (n=3204) Number	Prevalence
A	2376	74%	A1	2137	67%
			A2	350	11%
			A3	268	8%
L	972	30%	L1	694	22%
			L2	261	8%
			L4	218	7%
S	761	24%	S1	265	8%
			S3	374	12%
			S5	278	9%
U	871	27%	U1	819	26%
			U2	75	2%

Table 5.6 indicates that three in four students tested as A at some time in the study; about one in three students tested as L; one in four students tested as S; and one in four students tested as U. So, while the overall TFP (see Table 5.1) for S was 12% (1 in 8 tests), this alternative definition of prevalence reveals that 1 in 4 students tested as S at some stage in the study. Table 5.6 also indicates that 350 students (i.e. 1 in 9 students) tested as A2 at some stage in their test history, compared with an overall TFP of only 4% (1 in 25) for all 9862 tests. This gives a different slant on the importance of the codes to the researcher and the classroom teacher.

Note that while the TFP figures could be added, it is not possible to combine the SFP figures. For example, the number of students who were involved in tests coded as A2 or A3 (577 or 18%) is not simply the sum of the appropriate figures in Table 5.6 due to students contributing to counts for both codes. Similarly, the number of students who were involved in tests coded as L2 or L4 (450 or 14%) cannot be determined from Table 5.6. Further discussions of the number of students with various codes appear later this chapter when students' test histories are considered.

SFP by School Level

As discussed in section 5.1, for the purposes of calculating the SFP, all students were allocated to either the primary sample or the secondary sample according to the grade of their first test. (Note that it is not meaningful to determine SFP for grades or semesters as it is a longitudinal measure and requires several tests. Hence, the smallest sensible grain-size is the school level.) Table 4.5 indicates that 1079 students were first tested in primary school and 2125 students were first tested in secondary school. For the students in the secondary sample, the complete set of tests that they generated is considered, but for the students in the primary sample, only the tests that these students generated whilst at primary school are examined.

Hence, the creation of these two samples for the purpose of SFP results in all 3204 students being involved, but not all 9862 tests being considered. (In effect, 1414 tests are ignored by this procedure. Due to the project effect, these later tests are more likely to be A1 than the earlier tests by students, so this is gauged as a reasonable procedure.)

Table 5.7 provides the SFP of both the coarse and fine codes for the two school levels to examine the effect of age on the figures in Table 5.6. As before, additional columns provide the difference between the secondary and primary figures to emphasise the codes where there are significant changes. As expected, there are higher percentages of students from secondary school (compared with primary school) who are involved in tests coded as A or A1, with corresponding lower percentages of students (compared with primary school) involved in L or L1 tests. There are considerable differences, however, between the figures for TFP and SFP. For example, while the TFP of L is only 8% in the Secondary level (Table 5.2), the corresponding SFP is 16%.

Table 5.7: School level SFP of both coarse and fine codes

Coarse code	Primary SFP (n=1079)	Secondary SFP (n=2125)	Secondary-Primary	Fine code	Primary SFP (n=1079)	Secondary SFP (n=2125)	Secondary-Primary
A	53	78	+24	A1	45	70	+24
				A2	7	11	+5
				A3	6	8	+2
L	57	16	-41	L1	46	9	-37
				L2	11	5	-6
				L4	12	4	-8
S	25	22	-3	S1	10	7	-3
				S3	11	11	0
				S5	8	9	+1
U	26	25	-1	U1	26	23	-3
				U2	0	3	+3

SFP by School Group

As with the TFP, the figures for the SFP of each code and each school group are provided in two tables, Table 5.8 for the coarse codes and Table 5.9 for the fine codes. Note that the sum of the rows exceeds 100% as students may contribute to several columns.

Table 5.8: School level SFP of coarse codes by school group

School Level and School Group		Coarse codes			
		A	L	S	U
Primary	**(n=1079)**	**53**	**57**	**25**	**26**
SGA	(n=208)	18	81	13	23
SGB	(n=123)	44	48	42	16
SGC	(n=325)	54	68	27	30
SGD	(n=290)	68	49	29	29
SGE	(n=133)	82	21	14	26
Secondary	**(n=2125)**	**78**	**16**	**22**	**25**
SGA	(n=501)	65	24	24	28
SGB	(n=556)	76	10	19	18
SGC	(n=224)	82	29	22	35
SGD	(n=107)	79	15	27	27
SGE	(n=125)	94	5	19	16
SGF	(n=612)	84	12	22	26

Table 5.9: School level SFP of fine codes by school group

School Level and School Group		Fine codes										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=1079)**	**45**	**7**	**6**	**46**	**11**	**12**	**10**	**11**	**8**	**26**	**0**
SGA	(n=208)	11	5	4	73	9	10	2	6	6	23	0
SGB	(n=123)	37	7	3	39	6	7	13	22	12	16	0
SGC	(n=325)	45	7	9	54	16	19	12	12	5	30	1
SGD	(n=290)	61	9	6	37	11	9	13	12	7	29	0
SGE	(n=133)	75	5	3	11	8	7	5	5	5	26	0
Secondary	**(n=2125)**	**70**	**11**	**8**	**9**	**5**	**4**	**7**	**11**	**9**	**23**	**3**
SGA	(n=501)	58	10	6	15	7	5	5	13	9	26	3
SGB	(n=556)	68	8	6	5	4	2	7	7	8	17	2
SGC	(n=224)	76	13	11	17	10	7	6	13	9	33	4
SGD	(n=107)	74	13	11	7	6	4	7	19	9	26	4
SGE	(n=125)	88	16	6	1	3	2	9	9	10	13	3
SGF	(n=612)	75	14	11	5	4	3	7	12	8	24	3

As might be expected, SGA and SGE have similar features as before; SGA has fewer students involved in A and A1, and more students involved in L and L1. SGE has the reverse situation. A few other variations are now noted; over 40% of the students in primary schools in SGB are involved in testing as S, due to both S3 and S5. Similarly, nearly 30% of the students in secondary schools in SGD are involved in testing as S, mostly due to S3 (1 in 5 students test as S3).

All the calculations of SFP so far are underestimates of any theoretical *lifetime risk* for two reasons as explained below. Firstly, students who are less-tested dilute them; with less testing, students have less opportunity to contribute to several codes. Appendix 5 contains an analysis of the effect on the SFP of each code by restricting the sample of students to those with at least two tests, at least three tests and lastly, at least four tests. These final figures are presented in Chapter 9 as the improved estimates for the SFP of each code.

Secondly, the maximum number of tests being completed by a student in any one year is two. If a student moves through three or more ways of thinking in a year, then one or two tests cannot capture all the different states. What is shown in Chapter 6, however, is that for many students, the opposite problem exists; they remain in the same state for long periods of time.

5.3 Students' test histories

As mentioned in the introduction to this chapter, every student has an (ordered) list of codes indicating the tests completed by that student; these are referred to as students' test histories. Tracing the changes in the codes from one test to another should reveal changes in that student's thinking, although always with the possibility noted above that some changes between tests will have been missed. Section 5.2.3 provided the details of the SFP for each code, which indicated whether or not a particular code was present in each student's test history, but it did not consider, for example, the presence of two particular codes for the one student. Hence, to provide a fuller picture of which students had which codes, this section will provide a series of results from further analyses that focus on students' test histories.

5.3.1 Occurrences of coarse codes

Table 5.10 contains details of the number of students whose test histories contain various combinations of the four coarse codes. For example, the first row indicates that 412 students completed exactly one test which was coded A, while another 203 students completed exactly two tests which were both coded A, etc. The final entries in this row indicate that 1309 students (i.e. 41% of the sample) had test histories which contained only the coarse code A.

To determine, for example, the total number of students involved in A tests, we need to sum various rows of this table; the 1st row was mentioned (1309 students), plus the 5th, 6th and 7th rows, as well as four more rows that indicate students had other combinations of codes as well as A. Hence, all the figures in the student-focussed prevalence in the preceding section can be obtained by adding the appropriate rows in Table 5.10.

The shaded columns indicate impossible situations. For example, it is impossible for a student to have both A and L tests and yet to have only have one test in their test history. Note that there is no analogous table for the fine codes (like Table 5.10) as it would have 2^{11}-1 rows.

Table 5.10: Numbers of students with combinations of coarse codes

Details of coarse classifications		Total number of tests completed by each student							Total	
		1	2	3	4	5	6	7		
Exactly 1	A	412	203	248	229	172	40	5	1309	*41%*
	L	167	42	21	22	15	1	0	268	*8%*
	S	98	24	12	13	3	1	0	151	*5%*
	U	95	23	7	1	0	0	0	126	*4%*
Exactly 2*	AL	-	41	54	72	46	26	14	253	*8%*
	AS	-	39	47	77	33	12	4	212	*7%*
	AU	-	54	82	57	47	9	4	253	*8%*
	LS	-	15	16	19	10	3	0	63	*2%*
	LU	-	21	19	34	13	8	1	96	*3%*
	SU	-	26	20	15	11	2	0	74	*2%*
Exactly 3*	ALS	-	-	11	27	24	11	4	77	*2%*
	ALU	-	-	18	38	53	22	7	138	*4%*
	ASU	-	-	31	45	23	6	2	107	*3%*
	LSU	-	-	7	15	15	9	4	50	*2%*
Exactly 4*	ALSU	-	-	-	8	11	4	4	27	*1%*
Total		772	488	593	672	476	154	49	3204	*100%*

**in any order*

5.3.2 Occurrences of fine codes within coarse codes

There were 2376 students who tested as A at some point in this study (see Table 5.6 for SFP). Of these students, 1799 (76%) tested as A1 and were not involved in either A2 or A3 tests. The Venn diagram in Figure 5.3a indicates the combinations of fine A codes in which these 2376 students were involved. After "only A1", the next most common combination is A1 and A2 (189 students). An additional 24 students tested as A1, A2 and A3, so there are a total of 213 students with both A1 and A2 (189 + 24). While these figures are exact, they are heavily influenced by sampling; one of the reasons for testing as A1 but not A2 or A3 is that the student completed only one test, and only students who have completed 3 or more tests are eligible to have tested as all three fine A codes.

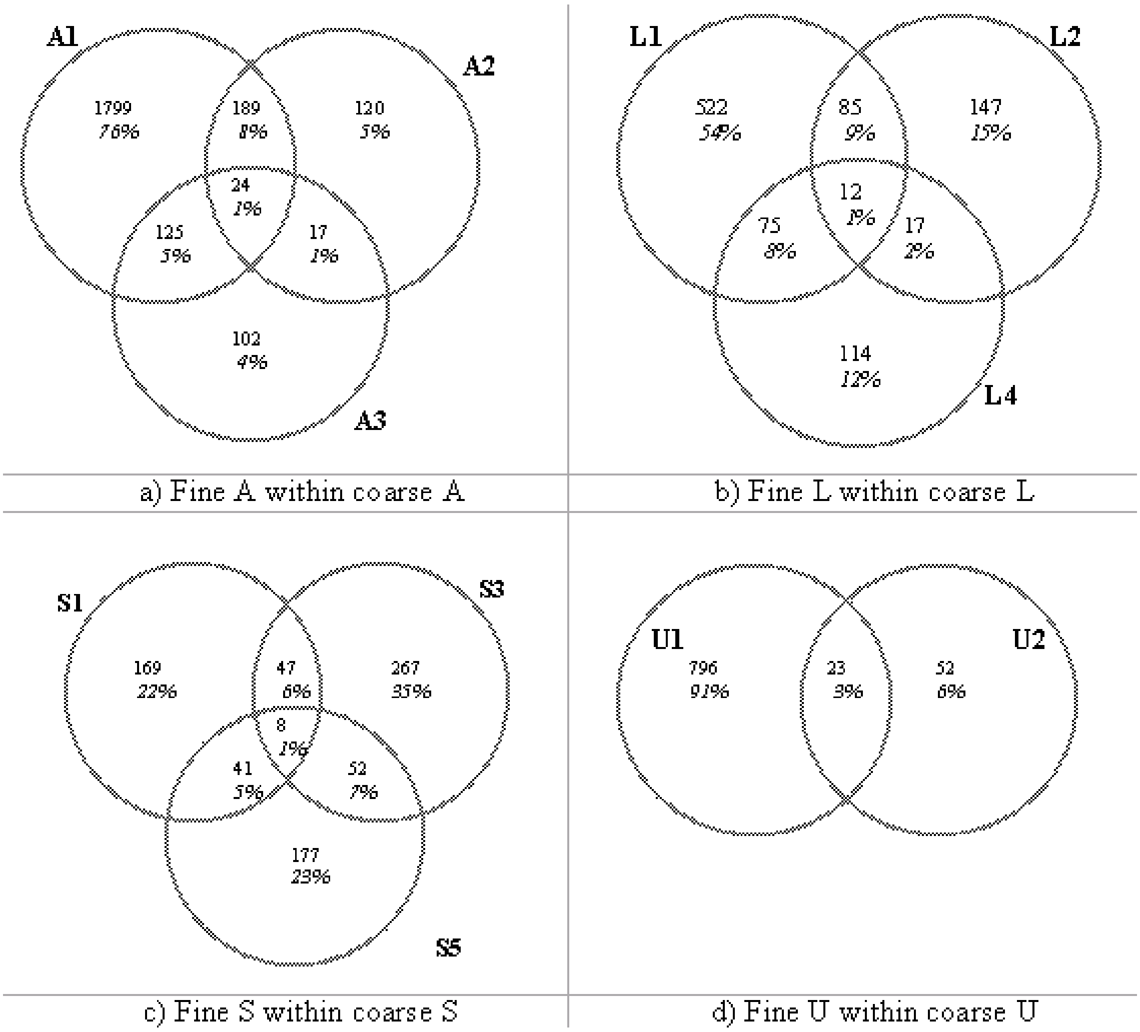


Figure 5.3: Venn Diagrams indicating occurrence of fine codes within coarse codes

Figures 5.3b, 5.3c and 5.3d contain similar details for the various combinations of fine codes within the coarse codes L, S and U, respectively. Note again the dominance by L1 within the L codes, and by U1 within the U codes, and that the situation is different within the S codes. While more of the S students are involved in S3, there are considerable numbers of students involved in S1 and/or S5.

While Figure 5.3 presents details of the combinations of fine codes within coarse codes, this initial analysis does not provide details of the order in which the codes appeared within student's test histories. These will be attended to in the next section.

Firstly, however, details of the test histories for the twelve students at the centre of the Venn diagram in Figure 5.3b are provided in Table 5.11. Similarly, Table 5.12 contains details of the eight students at the centre of the Venn diagram in Figure 5.3c. These will be referred to in later chapters.

Table 5.11: Details of twelve students with L1, L2 and L4 tests

ID	Grade 4	Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10	
100802011								L1			U1	L4	L2
300704112						L1	L4	L4	L2	L1			
310401041	L2		L1	U1	U1	L4	U1	U1					
310401061	L1		L4	L4		L2	L1	A1					
310403001	L1	L1	L4	L2	U1	A1							
310403044	U1	S3	L1	L2	L4	L2							
310403047	L4	L2	L1	S1	U1								
310403084	L4	L1	A1	L2	A1	A1							
410401088	L1	L1		L4	L1	L2	A1	A1					
410504030		L4	L1	L1	L2	U1							
410504041		L1	L4	A1	A1	L2							
510504018		L1	L4	L1	U1	L2							

Table 5.12: Details of eight students with S1, S3 and S5 tests

ID	Grade 4	Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10	
100702005						S1			S5		S3		
200703071							S1		S1		S3	S5	
210403026	L1		A1	S3	S5	S1							
300704069						U1	S1	S3	S1	S5			
500703003							S1	S5		S3	S3	U1	
500703030							S3	S5		S1	A2		
500703072							S1	S5		A2	S3		
600705119							S1	S1	S3	S5			

5.3.3 Students with pairs of codes

In this section, the numbers of students with two particular codes will be reported. In this analysis, each student's test history was considered to determine the presence or absence of pairs of codes. Note that these pairs of codes need not be consecutive.

For example, to determine the number of students with both A and L codes in their test histories, the appropriate rows of Table 5.10 can be added to give 495 students. This is repeated for all possible pairs of coarse codes; see Table 5.13. (Note that due to the symmetry of these tables, the lower triangles could be deleted, but this would prevent the inspection of full columns.)

Table 5.13: Number of students with any two particular coarse codes

Coarse code	Coarse Code			
	A	L	S	U
A		495	423	485
L	495		217	303
S	423	217		244
U	485	303	244	

Table 5.14: Number of students with any two particular fine codes

Fine code	Fine code										
	A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
A1		213	149	283	125	97	134	147	119	409	47
A2	213		41	38	23	20	39	40	33	69	10
A3	149	41		68	23	22	25	29	20	89	11
L1	283	38	68		97	87	62	72	46	219	4
L2	125	23	23	97		29	27	29	16	86	3
L4	97	20	22	87	29		16	24	18	65	3
S1	134	39	25	62	27	16		55	49	79	4
S3	147	40	29	72	29	24	55		60	120	12
S5	119	33	20	46	16	18	49	60		93	11
U1	409	69	89	219	86	65	79	120	93		23
U2	47	10	11	4	3	3	4	12	11	23	

Several points need to be made about these tables. Firstly, it is not appropriate to sum the entries in rows or columns as students are counted several times. For example, a student with a test history of [S,U,A] would have contributed three times to the above analysis; to the A and S count (423), the A and U count (485), and the S and U count (244).

Secondly, the 772 students with only one test are unable to make any contribution to these tables. This is not to say that every other student contributes to this table; a student with a test history of [L1,L4,L1] makes no contribution to Table 5.13, but will contribute once to the L1 and L4 count (87) in Table 5.14.

There are two analyses that are based on the data in these two tables and reported in later chapters. For example, Table 5.13 indicates that there are 495 students with both L and A tests *in any order*. In fact, the overwhelming majority of these students (462, or 93%) have their first L test before their first A test. Clearly then, L is a *younger/earlier* code than A. The order that the codes appear within a student's test history will be explored in Chapter 7 to determine *developmental trends*.

Another analysis that is based on this data is referred to as *association* and is reported in Chapter 9. In this analysis, the relative likelihood of a student who is known to be involved in one particular code, also being involved in other codes, is determined. Such information will shed more light upon the thinking behind the codes.

5.4 Transitions between codes on consecutive tests

In this section, pairs of consecutive tests from each student's test history are analysed to provide information on the *transitions* between codes. This provides some information about the codes and a useful set of results for future reference. This analysis is based on pairs of tests completed by a student. There are 772 students who have only completed one test and who therefore do not contribute to this analysis, while 488 students with exactly 2 tests [T_1,T_2] contribute just once. Students with exactly 3 tests [T_1,T_2,T_3] contribute twice. For example, a student with a test history of [L,S,A] would contribute two pairs of tests to this analysis; the first pair of tests in this case is LS then the second pair of tests is SA. Students with 7 tests contribute six test pairs.

Altogether, the 2432 students with at least 2 tests contribute 6658 pairs of tests in the following analysis. Note that the definition of consecutive tests is from the point of view of the student, rather than testing periods. So, for example, a student who tested as L in TP4, then missed being tested in TP5, and then tested as A in TP6, still contributes to the transitions corresponding to LA.

The following three analyses are conducted using the coarse codes, the fine codes and then a hybrid that combines both coarse and fine codes.

5.4.1 Coarse codes

The number of transitions from one coarse code to another on consecutive tests is provided in Table 5.15; the diagonal cells are shaded to indicate *no movement* by a student between the tests, as both tests have the same code. By far the largest entry in Table 5.15 is the 3408 test pairs that indicate AA. Note that it is expected that the test following an A test would also be A.

Table 5.15: Number of test pairs from coarse code on T_i to coarse code on T_{i+1}

Coarse code on T_i	Coarse code on T_{i+1}				Total
	A	L	S	U	
A	3408	48	87	203	3746
L	336	548	145	228	1257
S	303	76	319	149	847
U	381	78	113	236	808
Total	4428	750	664	816	6658

Table 5.16 shows the percentage distribution of next tests; Stacey and Steinle, (1999b) reported similar results, using a subset of this data. The row percentages for Table 5.16 were calculated from Table 5.15 as follows. In the last column of Table 5.15, for example, there are 3746 test pairs that have A as the first test; of these 3408 (or 91%) were A on the next test, (i.e. they stayed as A). This leaves 9% who were not A on the next test and will be referred to as *departures* from A. In particular, 48 (1%) went to L, 87 (2%) went to S and 203 (5%) went to U. This procedure is repeated for the remaining coarse codes and the results are provided in Table 5.16.

As before, the diagonal cells are shaded to indicate no movement between tests and these tend to be the largest entries within a column. The other large entries are in the first column, indicating movement to A. It is much less likely that students swap between L and S. Also note that the rate of U to A is relatively high compared with L and S, confirming Moloney and Stacey's (1997) finding, with a small sample, that the most likely students to move to expertise were previously unclassified.

Table 5.16: Distribution (%) of next test given coarse code of current test

Coarse code on current test T_i		Coarse code on next test T_{i+1} A	L	S	U
A	(n=3746)	91	1	2	5
L	(n=1257)	27	44	12	18
S	(n=847)	36	9	38	18
U	(n=808)	47	10	14	29
Overall	(n=6658)	67	11	10	12
non-A	*(n=2912)*	*35*	*24*	*20*	*21*

The last row in Table 5.16 provides summative information. The overall rate to A is 67%, but students who were already A inflate this. Hence, these students are disregarded in the calculation for the last row of this table. Overall, the rate of movement to A on the next test, by non-A students, is 35%.

Before considering the fine codes, another measure involving the coarse codes will be presented. It was mentioned above that as 91% of A tests are followed by A, then the *departure rate* from A is 9%. Similarly, the departure rates from L, S and U (56%, 62% and 71%, respectively) are determined as the complement of the diagonal entries in Table 5.16. An analogous measure of *arrival rate* can be determined from Table 5.15. Consider the 4428 test pairs that have T_{i+1}=A. Of these, 3408 test pairs involved A as the previous test, while 1020 did not. Hence the arrival rate to A is the proportion of test pairs which indicate a student has moved to A from a non-A test; here 1020/4428 = 23%. The arrival rates for L, S and U are similarly calculated to be 27%, 52% and 71%, respectively, and are listed in Table 3 in Appendix 7 for future reference.

Figure 5.4 provides a comparison of the arrival and departure rates for the four coarse codes. It indicates that code A has an arrival rate of about 25%, and a departure rate of about 10%, which is consistent with the earlier results that the TFP of A increases with grade. In contrast, the departure rate for L is about double the arrival rate, which is consistent with the earlier results that the TFP of L decreases with grade. For both S and U there is more of a balance between arrival and departure rates, which is consistent with the flatter distributions of these codes against grade.

The difference between the L and S codes is now clear; while they have similar departure rates (about 60%) S has twice the arrival rate of L. So, the proportion of new students to S (i.e. from A, L and U) is twice that for L (i.e. from A, S and U). That is, S ideas are “attractive” to students.

The code U has the highest turnover of these codes; a higher proportion of arrivals and departures compared with A, L and S. This is consistent with U indicating a student who is not an expert but who is not holding onto a particular misconception either. Such students appear to be in transition between other states.

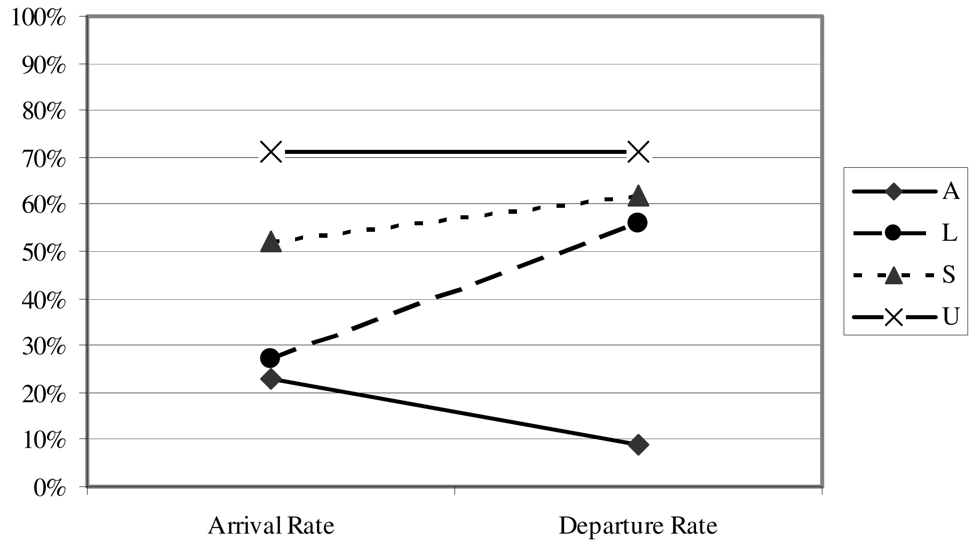


Figure 5.4: Comparison of arrival rate and departure rate for each coarse code

This concludes the analyses of the coarse codes generated from Table 5.15. Similar analyses of the fine codes will now follow.

5.4.2 Fine codes

The number of transitions from one fine code to another on consecutive tests is provided in Table 5.17, and the row percentages are provided in Table 5.18. Two different colours of shading have been applied to the diagonal cells; the darker shading indicates retesting in the same fine code, while the lighter shading has been applied to other cells which indicate movement within the same coarse code.

There is a strong tendency for students with an A1 test to remain the same on the next test (about 90%). This is stronger than the tendency to remain in any other code. The tendency to retest as L1 is almost 40%, and is the main reason for the 44% retest rate for L. Likewise, the retest rate of 33% for S3 is the main reason for the 38% retest rate for S. The tendency for students to retest in the same code is referred to as persistence and is examined in detail in Chapter 6.

Table 5.17: Number of test pairs from fine code on T_i to fine code on T_{i+1}

Fine code on T_i	Fine code on T_{i+1}											Total
	A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2	
A1	2930	87	47	7	19	1	18	16	19	112	23	3279
A2	147	53	17	2	5	2	7	14	6	21	6	280
A3	93	11	23	6	4	2	2	2	3	36	5	187
L1	128	18	33	322	64	53	25	40	23	146	1	853
L2	73	13	9	23	26	4	11	14	7	46	1	227
L4	50	3	9	29	17	10	7	9	9	34	0	177
S1	81	18	15	11	8	4	32	26	19	28	3	245
S3	86	10	8	21	10	5	27	127	23	62	6	385
S5	63	10	12	9	4	4						

Table 5.18: Distribution (%) of next test given fine code of current test

Fine code on current test T_i		Fine code on next test T_{i+1}										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
A1	(n=3279)	89	3	1	0	1	0	1	0	1	3	1
A2	(n=280)	53	19	6	1	2	1	3	5	2	8	2
A3	(n=187)	50	6	12	3	2	1	1	1	2	19	3
L1	(n=853)	15	2	4	38	8	6	3	5	3	17	0
L2	(n=227)	32	6	4	10	11	2	5	6	3	20	0
L4	(n=177)	28	2	5	16	10	6	4	5	5	19	0
S1	(n=245)	33	7	6	4	3	2	13	11	8	11	1
S3	(n=385)	22	3	2	5	3	1	7	33	6	16	2
S5	(n=217)	29	5	6	4	2	2	9	14	7	21	2
U1	(n=757)	37	4	5	4	3	3	5	7	3	28	1
U2	(n=51)	43	2	8	2	0	0	0	4	2	16	24
Overall	(n=6658)	59	4	3	7	3	2	3	5	2	11	1
Non-A1	*(n=3379)*	*30*	*5*	*5*	*14*	*5*	*3*	*5*	*9*	*4*	*19*	*1*

The last two rows of Table 5.18 provide summative information. The overall rate to A1 is 59%, but students who were already A1 inflate this. Hence, these students are disregarded in the calculation for the last row of Table 5.18. Overall, the rate of movement to A1 on the next test, by non-A1 students, is only 30%. Over an average period of 8.3 months, less than 1 in 3 of the non-expert students move to expertise.

As for the coarse codes, the arrival and departure rates have been calculated and are presented in Figure 5.5 (see Table 3 in Appendix 7 for details). Departures from A1, in particular, are significant as they indicate *regression*; see Chapter 8. Note the similarity between A1 in Figure 5.5 and A in Figure 5.4; similarly with L1 and L. Inspection of Figure 5.5 reveals that the next fine code with the lowest arrival and departure rates is S3, followed by U1. The remaining seven fine codes have reasonably similar, and high, arrival and departure rates; typically 75% to 95%. Note the positive slope on most of the lines in Figure 5.5; there are only three fine codes with negative slopes. The largest difference between the departure rate and arrival rate is for A1 (15%); the next largest is for U2 (6%) and then A3 (2%). The code U2 is investigated in Chapter 8.

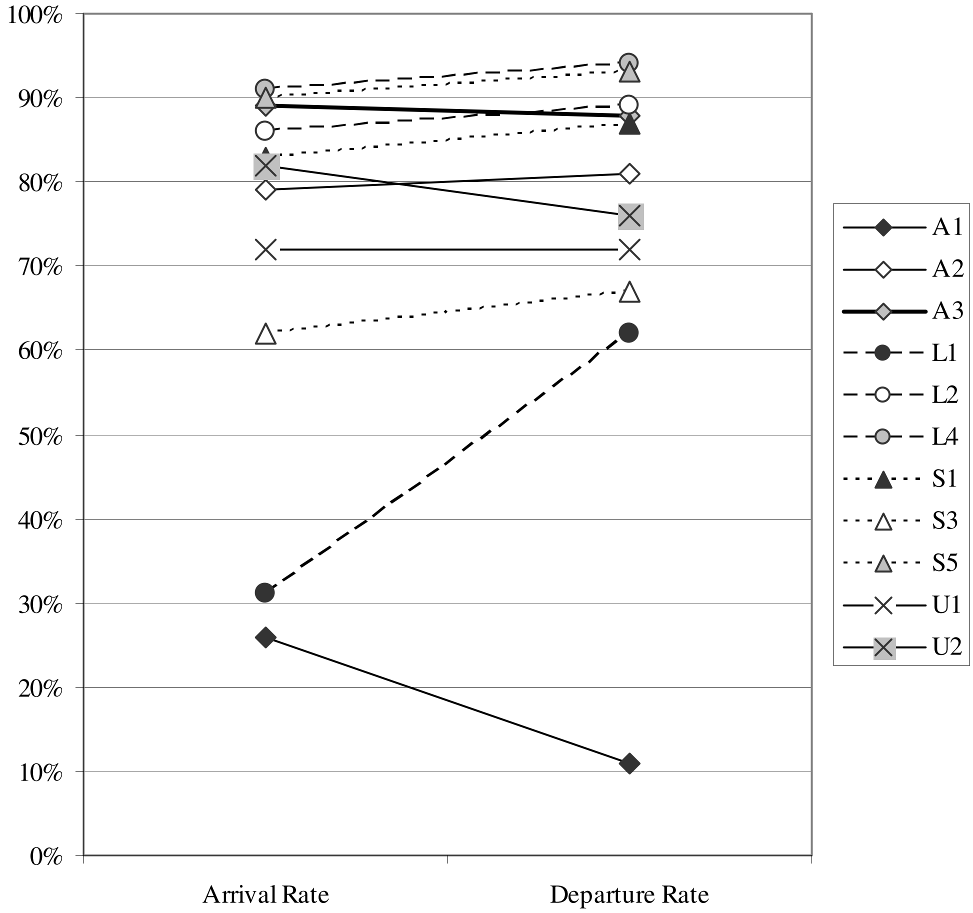


Figure 5.5: Comparison of arrival rate and departure rate for each fine code

Table 5.18 is the foundation table that illustrates three of the issues to be investigated with more refined methods later in this thesis. There are considerable variations on the rate of retesting in the same fine code, from a high of 89% for A1, to a low of 6% for L4; these are considered in Chapter 6, Persistence. Chapter 7 seeks to determine a hierarchy between the codes defined in terms of the relative movement to A1; for example, the code L1 has the lowest rate of movement to A1 (only 15%) and hence appears to be the most *primitive* code. Departures from A1 are referred to as *regressions* and are considered in Chapter 8.

Appendix 7 contains the details of the transitions between the coarse codes as well as the transitions between the fine codes, and will be referred to as required throughout this thesis.

5.4.3 Hybrid of coarse and fine codes

Figure 5.6 provides another way of summarising the results from Table 5.18. By combining the columns into the coarse codes and retaining the rows as the fine codes, this allows us to see the strong tendency to remain in the same coarse code and also allows comparison of the fine codes. Hybrids between the coarse and fine codes, such as this, will be used throughout this thesis as they offer a suitable balance between detail and clarity.

Figure 5.6 provides the distribution of the coarse code of the *next* test given the fine code of the current test (based on figures in Table 1 in Appendix 7). While bar graphs could have been used (as the sum of the distributions is 100%), the variations between the fine codes are easier to observe in these column graphs.

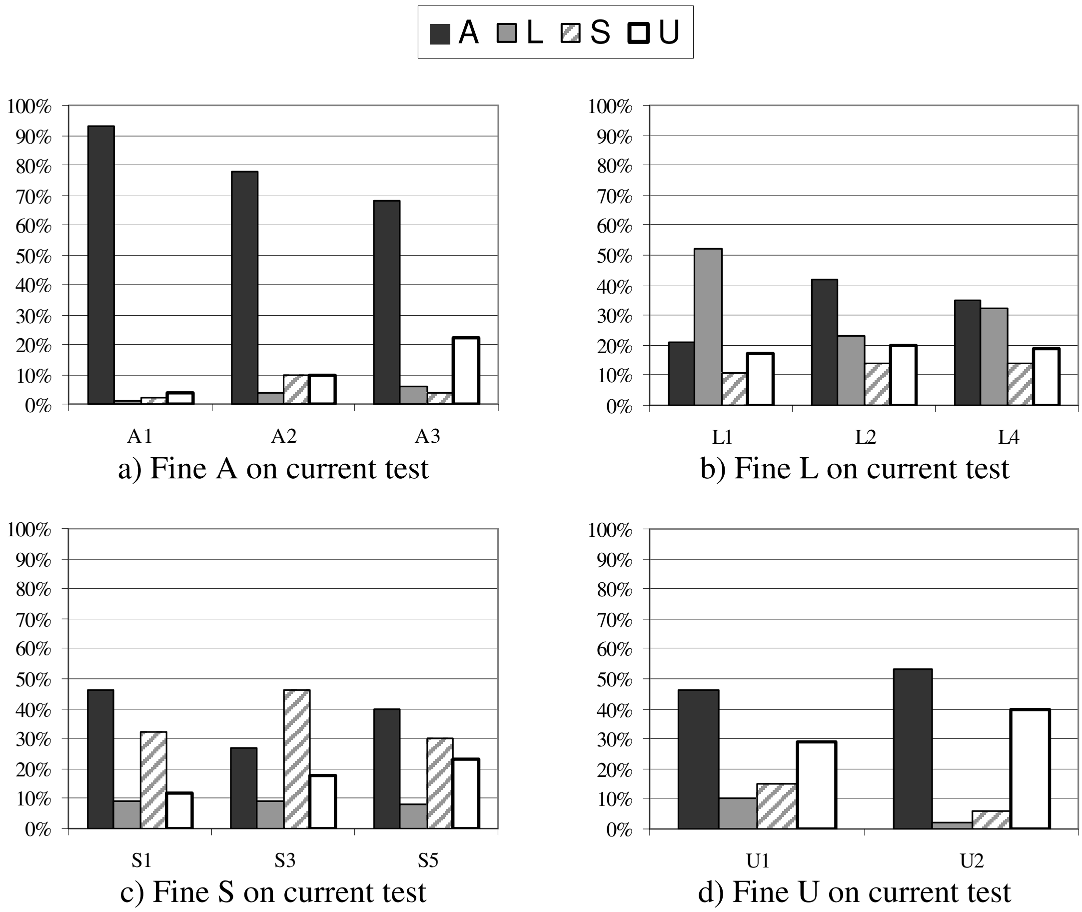
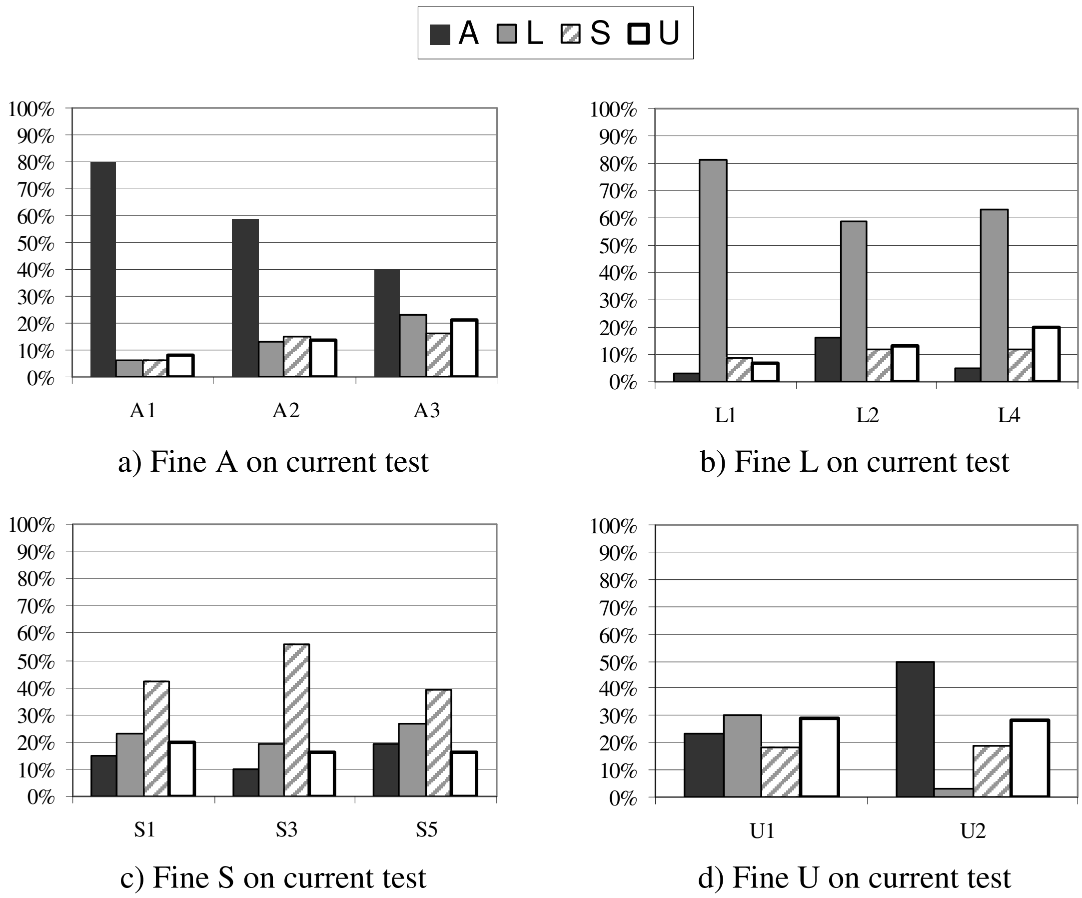


Figure 5.6: Percentage of coarse codes on next test, given fine code of current test

Figure 5.6a shows that there is an increased tendency for A3 tests (compared with A1 and A2 tests) to be followed by U tests. Furthermore, there are more A2 tests (compared with A1 and A3 tests) that are followed by an S test. Figure 5.6b shows the increased tendency for L1 (compared with L2 and L4) to retest in L on the next test, instead of moving to A. Similarly, Figure 5.6c shows the increased tendency for S3 (compared with S1 and S5) to retest in S on the next test, instead of moving to A. Figure 5.6d shows that about 50% of U2 tests are followed by an A test.

For completeness, Figure 5.7 provides the distribution of the coarse code of the *previous* test given the fine code of the current test (created from figures in Table 2 in Appendix 7). Figure 5.7a indicates that there is almost twice the chance (compared with an A1 or A2 test) that an A3 test was preceded by an L test. Figure 5.7b indicates that there is about three times the chance that (compared with L1 and L4) an L2 test was preceded by an A test. Figure 5.7d indicates that there is about twice the chance that (compared with U1) a U2 test was preceded by an A test.



a) Fine A on current test b) Fine L on current test

c) Fine S on current test d) Fine U on current test

Figure 5.7: Percentage of coarse codes on previous test, given fine code of current test

Various pieces of evidence such as this will be summarised in Chapter 9 when the links between the codes and the ways of thinking are discussed.

5.5 Conclusion

Both the test-focussed prevalence (TFP) and the student-focussed prevalence (SFP) of the coarse and fine codes have been determined and the variations with students of different ages and from different school groups have been reported. Not surprisingly, the levels of expertise increase with age. What is a surprise, however, is that by Grade 10, one in three students are still not experts. This is even more surprising given that there are two, apparently simple, expert algorithms available for this task.

The variations between the school groups in the levels of expertise are also surprising; for example, SGE has 90% of students in Grade 9 testing as experts compared with only 60% of students of this age in SGA. Also surprising is that by Grade 10, one in ten students are completing tests which are coded as either A2 or A3. This indicates that while they are able to make correct comparisons on the ten core items in the test, they are unable to consistently choose correctly on items that involve zeros or on items where the digits in the tenths and hundredths columns are identical. Their strategies give correct answers on particular, possibly more *commonly met*, comparisons, so their inadequate understanding of decimal notation is often not revealed in the classroom. These students will be investigated further in Chapter 8.

The test-focussed prevalence of students exhibiting L behaviour decreases with age. It is more alarming, however, that the numbers of students exhibiting S behaviour remains fairly constant (10-15% of most grades).

The test-focussed prevalence of the coarse code A is dominated by A1 (about 90%); a similar situation holds for L with dominance by L1 (about 70%) and for U (dominance by U1). Within the coarse code S, however, a different pattern emerges. Overall, S3 contributes only half of S, with S1 and S5 contributing equally to the other half.

The student-focussed prevalence (SFP) of the codes provides a different perspective on the significance of the various codes to both researchers and teachers. For example, A2 is involved in 3% of the tests completed by students in secondary school, yet more than 10% of students in secondary school completed tests coded as A2 *at some time in the study*. The figures obtained for the SFP of the codes underestimate the true values as some students in the sample have only completed one test. Improved estimates will be provided in Chapter 9.

A positive project effect of about 10% has been detected, as hypothesised. This would be due to several reasons as discussed in section 4.2.5 including; students doing the same test on several occasions, and interested teachers may have used the feedback provided by the researchers to inform their subsequent teaching. This improved performance needs to be considered in the remainder of this thesis as further investigation into a student's test history is undertaken. It will also mean that only the first tests will be considered when the improved estimates of the prevalence of expertise are being determined in Chapter 9.

Considering the movement of students in and out of the codes provides broad comparisons between the coarse codes. Focussing on the current code, and identifying the *arrivals* (proportion of students moving into that code from another code on the previous test) and similarly the *departures* (proportion of students moving out of that code to another code on the next test) indicates the main differences between the codes. The code A is growing in prevalence as the arrival rate exceeds the departure rate, in contrast to L which has a higher departure rate than arrival rate and is therefore in decline. While S has a similar departure rate to L, the arrival rate is double that of L; that is, twice as many students are moving into S (from A, L or U) than into L (from A, S or U). The reason the code S has a fairly constant prevalence is the balance between the arrival rate and departure rate. The code U similarly has a balance between arrivals and departures, but the rates are higher than for all the other codes; hence students move in and out of U more than the other codes. This supports the view that students who complete tests coded as U are in transition.

The analysis of consecutive test pairs has set the scene for the following three chapters. Chapter 6 has an investigation into *persistence* that is stimulated by the finding that, for many non-expert codes, there is a high rate of retesting in the same code on the next test. Chapter 7 has in investigation into *hierarchy* of the codes that will be determined by examining the relative likelihood of movement to expertise on the next test as well as developmental trends. Chapter 8 contains an investigation into *regression* that is stimulated by the finding that about one in ten A1 tests is followed by a non-A1 test. The differences and similarities between the various codes will be summarised in Chapter 9.

CHAPTER 6 PERSISTENCE

6.1 Introduction and terminology

How long do students retain the same misconception? Are there some misconceptions that are harder for students to overcome than others? Using a small sample, Moloney and Stacey (1997) found that almost all students retested in the same category one year later; Helme and Stacey (2000a) confirmed this result.

Is there evidence within this dataset of students persisting with the same misconceptions over long periods of time? If a student uses the same way of thinking for several consecutive tests, then there will be a *sequence* of equal test codes in their test history. Are all codes equally likely to occur as sequences? This would indicate that these codes represent misconceptions which are harder for students to overcome.

Stacey and Steinle (1999b) used a subset of this data for a preliminary investigation into transitions between the coarse classifications. They found that students exhibiting either L behaviour or S behaviour on the test had a tendency to retest the same on a later test. In particular, after a period of about two years, 20% of L students were completing additional L tests; similarly S students were completing additional S tests. This analysis will be extended in this chapter using the fine codes.

Terminology

Students who retest in the same code on consecutive tests will have a *sequence* of codes in their test history. For example, a student with a test history of [L1,S3,S3,A1] has an S3 sequence of length 2. Another student may have two occurrences of the code S3 in their test history, for example [S3,U1,L2,S3,A1], but these are not consecutive occurrences; they are *scattered* occurrences.

Persistence refers to the tendency to retest in the same code. *Persistence over one semester* is the proportion of students who retest in the same code on the next test, which is usually in the next semester (on average, 8 months later). *Persistence* can also be determined over longer periods.

Section 6.2 provides a preliminary analysis of which codes appear repeatedly in students' test histories. Section 6.3 investigates sequences of codes within students' test histories. Section 6.4 further probes the variations in persistence with grade and school group, while section 6.5 considers persistence over various semesters.

Various graphs are provided in this chapter to illustrate the issue of persistence. The sizes of the samples, on which the calculations are based, are provided in Tables 31 to 34 in Appendix 8.

6.2 Average number of occurrences

An interesting and simple calculation can be made with figures from Chapter 5. The total number of tests allocated a particular code was given in Table 5.1 while the corresponding number of students involved in a particular code was given in Table 5.6. The ratio between these figures provides a preliminary indication of which codes occur repeatedly for some students. For example, 1596 tests were allocated L and there were 972 students with at least one L test, so these students had an average of 1.64 L tests each. Table 6.1 contains the average number of occurrences for involved students for all the codes; note that, by definition, the minimum value for this average is 1.

Table 6.1: Average number of occurrences of codes for involved students

Coarse code	Average number of occurrences for involved* students	Fine code	Average number of occurrences for involved* students
A	2.50	A1	2.44
		A2	1.19
		A3	1.11
L	1.64	L1	1.54
		L2	1.14
		L4	1.06
S	1.50	S1	1.19
		S3	1.41
		S5	1.08
U	1.37	U1	1.35
		U2	1.17

** An involved student is one who has at least one occurrence of this code*

It is not surprising that code A is more likely to occur again in later tests by a student, compared with L, S or U. In fact, Table 5.16 showed that over 90% of A tests are followed by another A test. Also not surprisingly, the fine code most likely to be involved in later tests (given a student has one such occurrence) is A1. What is surprising, however, are the high values for L1 (1.54), S3 (1.41) and U1 (1.35). While this is only a preliminary analysis (as it does not take the actual order of tests into account) it does provide evidence that the codes L1, S3 and U1 are different to the remaining codes. These codes will be further investigated in this chapter.

6.3 Analysis of sequences

This section provides details of sequences of codes that occur in students' test histories. It is mostly descriptive and, as will be explained, it provides a best-case scenario.

6.3.1 Grades involved in L and S sequences

Firstly consider two students with sequences of length six. There is only one student in the study with a sequence of six L codes; student 310403080 with a test history of [L1,L2,L1,L1,L2,L1] completed tests in six consecutive testing periods (from Semester 2 of Grade 4 to Semester 1 of Grade 7). Similarly, there is only one student in the study with a sequence of six S codes; student 510603017 with a test history of [S2,S3,S3,S3,S3,S3] completed tests in six consecutive testing periods (from Semester 2 of Grade 6 to Semester 1 of Grade 9). The code S2 indicates that this student completed DCT1, which was less able to diagnose S3, and so it is possible that this student may have always been using one of the ways of thinking that is usually coded as S3.

How typical are these two students? The first student comes from SGC and has a sequence of L codes that ends in the Grade 7 (when the student left the study). The second student comes from SGE and the sequence of S codes is from tests from Grades 6 to 9 and then also left the study. Can we conclude that sequences of L codes are a primary school phenomenon and that sequences of S codes are a secondary school phenomenon? Are students with sequences (i.e. consecutive tests with the same code) to be found in every school group or just some?

To answer the first question, Table 6.2 contains details of the first and last L test (in a sequence of L tests) for the 972 students who tested as L at some time in the study, and Table 6.3 contains similar details for the 761 students involved in S. The first row of Table 6.2, for example, indicates that 244 students first tested as L in Grade 4 and that 5 of these students, including the student 310403080 mentioned above, (indicated by *a*) continued to retest as L until Grade 7. The third row of Table 6.3 indicates that there were 118 students who first tested as S in Grade 6 and then only one of those, student 510603017 mentioned above, (indicated by *b*) continued retesting as S until Grade 9.

Table 6.2: Numbers of students with L sequences starting and finishing in particular grades

Grade of first L in sequence	Grade of last L in sequence							Total
	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10	
Gr 4	133	59	47	5[a]				244
Gr 5		164	63	21	2			250
Gr 6			100	12	6	1		119
Gr 7				173	28	16	1	218
Gr 8					72	9	2	83
Gr 9						27	6	33
Gr 10							25	25
Total	133	224	209	211	108	53	34	972
Last test	*40*	*35*	*74*	*57*	*38*	*41*	*30*	*315*

a: Includes student 310403080 [L1,L2,L1,L1,L2,L1]

Table 6.3: Numbers of students with S sequences starting and finishing in particular grades

Grade of first S in sequence	Grade of last S in sequence							Total
	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10	
Gr 4	17	6	3					26
Gr 5		97	24	4				125
Gr 6			99	16	2	1[b]		118
Gr 7				155	42	32	3	232
Gr 8					109	26	8	143
Gr 9						60	11	71
Gr 10							46	46
Total	17	103	126	175	153	119	68	761
Last test	*2*	*15*	*30*	*31*	*56*	*75*	*58*	*267*

b: Student 510603017 [S2,S3,S3,S3,S3,S3]

Note that, while L has been determined as a code more involved in primary school than in secondary school, a total of 406 students (42% of the 972 students involved in testing as L) completed their L sequence in secondary school. Hence, there is clear evidence that there are students in secondary school who repeatedly test as L and so this is not just a concern for primary teachers. Of the 761 students involved in S tests, 515 students (68%) completed their S sequences in secondary school, so it appears that students repeatedly testing as S is more a secondary than primary phenomenon (see Table 6.3).

Sampling heavily influences both tables. If all students were involved in testing from Grade 4 to 10, then the empty cells would be significant as they would indicate that students were involved in tests other than L (Table 6.2) or S (Table 6.3). In this study, however, students joined the study and left the study in various grades; for example see Table 4.6 for students' first and last tests.

The last rows of Tables 6.2 and 6.3 indicate the number of students who stopped their sequence, not because they moved to another code, but because it was their last test. A total of 315 students (32%) stopped testing L for this reason and 267 (35%) stopped testing as S. Hence, these sequences reveal a best-case scenario; continuing the testing program would probably have created longer sequences into older grades.

6.3.2 Lengths of sequences

The above discussion focussed on the grades involved in the first and last tests of L and S sequences. The *lengths* of the sequences for all the codes are now presented. Table 6.4 contains the number of students with sequences of varying lengths; note that many students will contribute to several rows (e.g. a student with a test history of [L1,A3,L1,L1,S3,S3,A3] would contribute two sequences of length 2). In particular, the number of involved students in each code comes from Table 5.6 (Student-Focussed Prevalence). Appendix 8 contains details which are summarised below.

The shaded cells in Table 6.4 indicate students who have a sequence of two or more non-A1 codes within their test history. In particular, there are 14 students (9+3+2) with sequences of five non-A1 codes (i.e. L1, S3 and U1). These students will be discussed below.

The average length of the sequences has been included in the last column of Table 6.4. Ranking the coarse codes from the lowest to the highest average gives U, S, L then A. It is reasonable that code A would have the longest sequences as students who test as

A on one test would be expected to retest as A on all later tests. The ranking of the fine codes is most easily seen in Table 6.5.

Table 6.4: Number of students with a sequence of length n for each code

Code	Length of the longest sequence							Number students involved	Average sequence length
	1	2	3	4	5	6	7		
A	870	490	447	318	198	48	5	2376	2.43
L	646	190	73	42	20	1		972	1.56
S	555	132	47	23	3	1		761	1.41
U	686	147	29	7	2			871	1.27
A1	816	446	398	272	163	41	1	2137	2.37
A2	311	27	10	2				350	1.15
A3	250	14	3	1				268	1.09
L1	485	133	48	19	9			694	1.46
L2	237	22	2					261	1.10
L4	208	10	0					218	1.05
S1	236	26	3					265	1.12
S3	292	56	13	10	3			374	1.33
S5	265	12	0	1				278	1.05
U1	656	130	25	6	2			819	1.25
U2	65	8	2					75	1.16

Table 6.5: Ranking of fine codes according to average sequence length

Fine Code	Number of students involved	Average sequence length
A1	2137	2.37
L1	694	1.46
S3	374	1.33
U1	819	1.25
U2	75	1.16
A2	350	1.15
S1	265	1.12
L2	261	1.10
A3	268	1.09
L4	218	1.05
S5	278	1.05

Note the three codes with the lowest average sequence length are S5, L4 and A3; the *fine unclassifieds*. In fact, these three codes do not indicate a known way of thinking, just a tendency towards A, L and S. Their position at the bottom of this ordering confirms that students tend to stay less in these codes than the other fine codes. This is despite the fact that the patterns of responses for these codes are less specified than for the other fine codes (see Table 3.3); if students responded randomly to items, repeats would occur more frequently.

To answer the second question posed earlier: *Are students with sequences of non-A1 codes found in all school groups?* the students in the shaded cells in Table 6.4 are combined and then split by school group; see Table 6.6 where the rows are ranked by the number of students in the first column.

While L1 is the code most likely to be involved in sequences, note that there are only two students in SGF with sequences of this code, probably as SGF is composed of older students. SGF has, however, about half of the students who are involved in A2 sequences.

Table 6.6: Number of students with sequences of two or more non-A1 codes by school group

Code	Number of students	School Groups					
		SGA	SGB	SGC	SGD	SGE	SGF
L1	209	76	16	87	24	4	2
U1	163	31	14	51	27	14	26
S3	82	8	14	19	20	10	11
A2	39	3	2	10	6	2	16
S1	29	4	3	11	6	3	2
L2	24	5	3	10	5	0	1
A3	18	1	4	8	4	0	1
S5	13	5	1	4	0	0	3
L4	10	1	0	7	2	0	0
U2	10	2	2	1	1	1	3

The data in Tables 6.4 and 6.6 indicates that some students retain the same misconception for long periods, as they move from one grade to another over their years at school and that this occurs in every school group. Due to the different number of students in each school group and the different testing regimes, (recall that the average number of tests per student in SGA, SGB and SGF was less than 3 and that in SGC, SGD and SGE, the average was above 4), no reliable quantitative evidence on school group differences can be obtained from Table 6.6. To determine how often students in each school group do have two consecutive equal tests, the following procedure was used.

A reduced sample of students was created, by excluding those students who only tested as A1, as well as those whose first non-A1 test was on their last test. This procedure was adopted as these students could not possibly have a pair of equal consecutive non-A1 tests in their test history. Of the 1637 students from this reduced sample, 560 (34%) were involved in retesting in the same non-A1 code on at least two consecutive tests.

The last column of Table 6.7 contains these details, and a breakdown of these 560 students by school group verifies that all school groups are involved to some extent. While the overall figure is one in three students, this varied from a maximum of one in two students in SGC to a low of one in four students in SGB and SGF. Note that a school group, like SGC, which has more tests per student, leads to more opportunities for this phenomenon to be detected. Hence, the overall rate of one in three students with at least two consecutive non-A1 tests will be an underestimate of this phenomenon due to the contribution of school groups with less testing.

Table 6.7: Numbers (and percentages) of students with pairs of equal non-A1 tests by school group

School Group		Number of pairs of equal non-A1 tests				Number students involved	Potential students*	%
		1	2	3	4			
Less-tested	SGA	87	28	10	4	129	355	36%
	SGB	48	9	2		59	249	24%
	SGF	41	18	6		65	271	24%
More-tested	SGC	112	44	22	9	187	387	48%
	SGD	58	18	9	2	87	265	33%
	SGE	27	5	0	1	33	110	30%
Overall		373	122	49	16	560	1637	34%

** students with their first non-A1 before their last test*

6.3.3 Sequences and scattered occurrences

It was found above that, after A1, the codes L1, S3 and U1 had relatively high average number of occurrences per student and that many students had these occurrences in sequences. How many occurrences of the codes were scattered throughout students' test histories, rather than being involved in sequences? These scattered occurrences would not have contributed to the above analysis, yet would provide information about students returning to the same way of thinking later in their test history.

Firstly consider some unusual students. Student 410401003 has a test history of [L1,A3,L2,A1,L2,A1] and therefore a total of 3 occurrences of L; but as these are not consecutive, the longest L sequence has length 1. Similarly, student 410401065 has a test history of [L1,S1,L1,S1,L1,U1,S1] and has 3 scattered occurrences of L; again the longest L sequence has length 1. These two students will now be placed in context, by considering all students with L tests.

Table 6.8 provides details of the 972 students who tested L at some stage. Each student is placed in a cell according to both the longest sequence of L and the total number of L's in their test history. The diagonal entries indicate students whose only occurrences of the code L are within a sequence, while the shaded cells indicate students who have scattered occurrences (i.e. they test as L, then move to another code and then return to L later). For example, the first row indicates that 646 students have longest L sequences of only 1; within these, 604 have no additional L tests, while 40 have just one scattered L test and the two students described above have two additional scattered L tests.

(Note: The procedure used to identify "the longest sequence" of a particular code in a student's test history resulted in the identification of the first sequence if there were two sequences of equal length. Hence, details in the rest of this chapter actually refer to the "first longest sequence" if there are several sequences of equal length.)

Most students are on the main diagonal (899 out of 972 or 92.5%) indicating that they have no additional occurrences of the L other than in the identified sequence. Hence, after they leave L, they do not return.

Table 6.8: Comparison of longest sequence and total number of occurrences of L in test histories

Longest L sequence	Total number of L's in the test history						Total number of students
	1	2	3	4	5	6	
L	604	40	2*				646
LL		166	22	2			190
LLL			69	4			73
LLLL				39	3		42
LLLLL					20		20
LLLLLL						1	1
Total	604	206	93	45	23	1	972

**Students 410401003 and 410401065*

The shaded cells indicate the 73 students (7.5%) who have scattered occurrences of L. Note that all of these figures are dependent on the sampling; students with only one or two tests cannot possibly move out of L and then back in, as this requires a minimum of three tests. Hence, this table illustrates the actual data but all numbers are underestimates of the severity of the problem. Similar calculations, for the percentage of students with scattered occurrences, were undertaken for A (6.1%), S (7.9%) and U (9.6%); see Appendix 8 for details. Hence, the code U is involved in more scattered occurrences than A, L or S. This is a reasonable result as it confirms that students move in and out of U; see section 5.4.1.

The percentages of students with scattered occurrences of the fine codes were also determined; Table 6.9 contains these details with the fine codes being ranked according to this measure. The fine codes vary from a maximum of 9.2% for U1, to a minimum of 1.3% for U2. After U1, comes a group of four codes with proportions between 6% and 7% (A1, L1, S1 and S3), then A2 and L2 with about 4% and then the three *fine unclassifieds* (S5, A3 and L4) with proportions of 2% to 3%.

Hence, this confirms earlier indications about the coarse code U being the most transient code; see Figure 5.4. While the number of occurrences of L1, S3 and U1 are all quite high, U1 has more scattered occurrences within student's test histories. So, in comparison to the codes S3 and L1, more students test as U1, then leave and then return to U1 later. This is reasonable, as U1 does not represent a particular way of thinking as much as a collection of students who do not fit any of the other ways of thinking. This will be further discussed in Chapter 9.

Table 6.9: Ranking of fine codes according to percentage of students with scattered occurrences

Code	Number of students involved	Students with scattered occurrences	
		Number	Percent
U1	819	75	9.2
L1	694	48	6.9
A1	2137	146	6.8
S1	265	18	6.8
S3	374	25	6.7
A2	350	15	4.3
L2	261	10	3.8
S5	278	8	2.9
A3	268	7	2.6
L4	218	4	1.8
U2	75	1	1.3

6.4 Persistence over one semester

In Chapter 5, transitions between pairs of (consecutive) tests were considered; these tests were determined to be, on average, 8 months apart. The rate of retesting in the same code between consecutive tests will be referred to as the *persistence over one semester*. We would predict that students in secondary schools are less likely (than their counterparts in primary schools) to persist with an incorrect way of thinking; variations in persistence with age will now be investigated.

We would also predict that a student who makes errors on equal length decimal comparisons would not persist with this way of thinking. In particular, a student using either *reciprocal thinking* or *negative thinking* would choose 0.3 > 0.4 and is expected to complete a test coded as S3. Such a student would be expected to make frequent errors in the classroom as many contexts provide decimals of equal length. It would therefore seem reasonable that such students would become aware that they have a problem, question the appropriateness of their current thinking and then move out of this code quite quickly.

The main diagonal of Table 5.16 reveals that, over one semester, students' persistence in A is 91%. Of the remaining coarse codes, L has the next highest persistence (44%), then S (38%) and U (29%). The relevant values for the fine codes come from the main diagonal of Table 5.18; they vary from a maximum of 89% for A1, to a minimum of 6% for L4. These are most easily seen in Table 6.10, which provides a ranking of fine codes according to persistence over one semester. After A1, the codes with the next highest persistence are L1, S3 and U1. Three of the four codes with the lowest persistence are the *fine unclassifieds* (L4, S5 and A3). In particular, L1 and S3 stand out as being codes that students have trouble leaving. As the L1 students tend to make correct comparisons of decimals with equal length, this result is expected, but the S3 code is allocated to tests that include 0.3 > 0.4, so it is unexpected that students are persisting in S3 to this extent.

Table 6.10: Ranking of fine codes according to persistence over one semester

Fine code	Number of suitable consecutive test pairs	Persistence (%) over one semester
A1	n=3279	89
L1	n=853	38
S3	n=385	33
U1	n=757	28
U2	n=51	24
A2	n=280	19
S1	n=245	13
A3	n=187	12
L2	n=227	11
S5	n=217	7
L4	n=177	6

A further investigation into persistence is now undertaken by considering the school level of the tests.

6.4.1 Variations with age

An initial investigation of the variations with age on these persistence values is undertaken by grouping the test pairs according to whether the first test in the pair was completed in primary or secondary school. Table 6.11 contains the primary and secondary level persistence over one semester as well as the difference. It seems reasonable to expect that younger students are more likely to persist with a misconception than secondary students, so the entries in the last column are expected to be negative.

The two coarse codes with large differences in the last column of Table 6.11 are L and S. Compared with students in primary school, students who test L in secondary school are *less likely* to retest as L (a predicted result). The opposite is true, however, for students testing as S; older students are *more likely* to retest as S. (As will be discussed below, these results are found in many of the school groups, they are not isolated to one unusual group of students.)

While many of the entries for the fine codes in the last column of Table 6.11 are negative or close to zero, the fine codes which give the unexpected (positive) difference are A2, S3 and, to a smaller extent, S5.

Table 6.11: Persistence (%) over one semester for the secondary and primary samples

Codes	Primary sample		Secondary sample		Secondary-Primary
A	90	(n=801)	91	(n=2945)	+2
L	48	(n=867)	35	(n=390)	-13
S	30	(n=311)	42	(n=536)	+12
U	27	(n=322)	31	(n=486)	+4
A1	90	(n=700)	89	(n=2579)	0
A2	7	(n=57)	22	(n=223)	+15
A3	16	(n=44)	11	(n=143)	-5
L1	39	(n=633)	33	(n=220)	-6
L2	13	(n=122)	10	(n=105)	-4
L4	6	(n=112)	5	(n=65)	-2
S1	15	(n=107)	12	(n=138)	-3
S3	26	(n=129)	37	(n=256)	+11
S5	4	(n=75)	8	(n=142)	+4
U1	27	(n=321)	28	(n=436)	+1
U2	0	(n=1)	24	(n=50)	*

* *small sample size*

Figure 6.1 contains the persistence in A2, L1, S3 and S5, over one semester, given the current grade; U1 is a flat distribution 25% to 30% at each grade and so is not graphed. (These values are contained in the shaded columns of Tables 12, 18, 28 and 30 in Appendix 7.) Note the decreasing trend for L1; due to the small number of L1 students in Grades 9 and 10 (n=14) the graph is not continued past Grade 8. There is a clear increasing trend for the persistence of students in A2, (as reported by Steinle and

Stacey, 2002) and while S3 has an increasing trend up to Grade 8, this drops in Grades 9 and 10. Although the persistence in S5 is lower than the other codes, the higher persistence in Grades 8 to 10, compared with Grades 4 to 7, is noticeable.

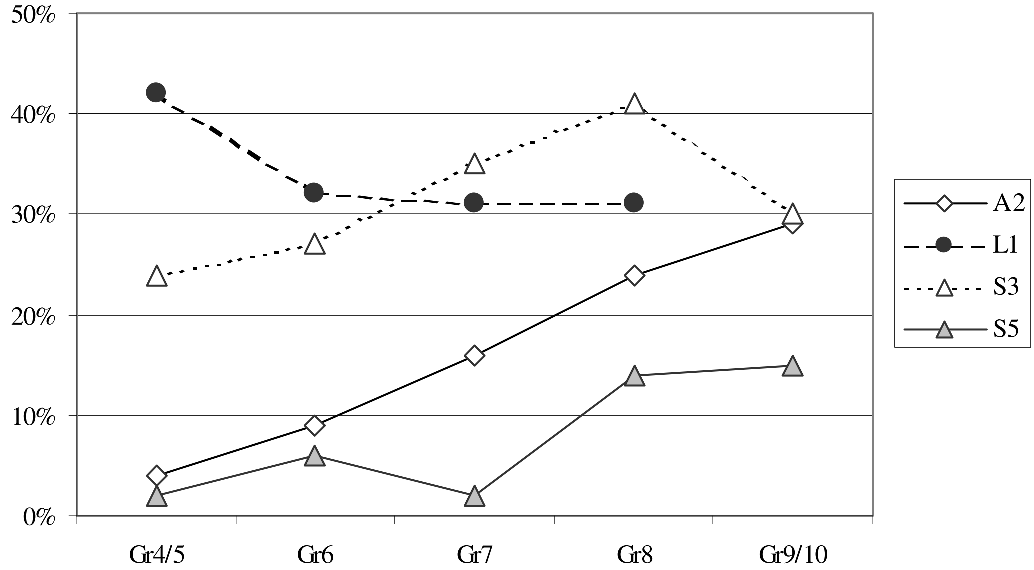


Figure 6.1: Persistence in A2, L1, S3 and S5 over one semester by grade of current test

In summary, as students get older, they are more likely to persist in A2 and S3. For A2 students it may be the case that they are becoming very "comfortable" with their way of thinking; possibly frequently rounding results of calculations to two decimal places may be reinforcing their beliefs. For S3 students, the high persistence is a surprise as these students consider 0.3 to be greater than 0.4 and this would be expected to be providing them with frequent cognitive conflict. A possible explanation is that exposure to negative numbers and also negative powers in Grades 7 and 8 is reinforcing their misunderstandings. Students with S5 also have an increasing persistence in older grades. Interference from new teaching was one of the causes of misconceptions listed in section 2.2.

To examine the S codes more fully, an alternative analysis involves a hybrid of coarse and fine codes. Recall that the entries in several columns of Table 5.18 were combined to create Figure 5.6. In particular, Figure 5.6c contains the transitions from fine S to coarse S codes; about 30% of the tests following an S1 (or S5) test were coarse S, and about 45% of the tests following an S3 test were coarse S codes. These results are now split by the grade of the current test; see Figure 6.2. There are definite trends in all three fine S codes- a peak in either Grade 7 (for S1 tests) or Grade 8 (for both S3 and S5 tests).

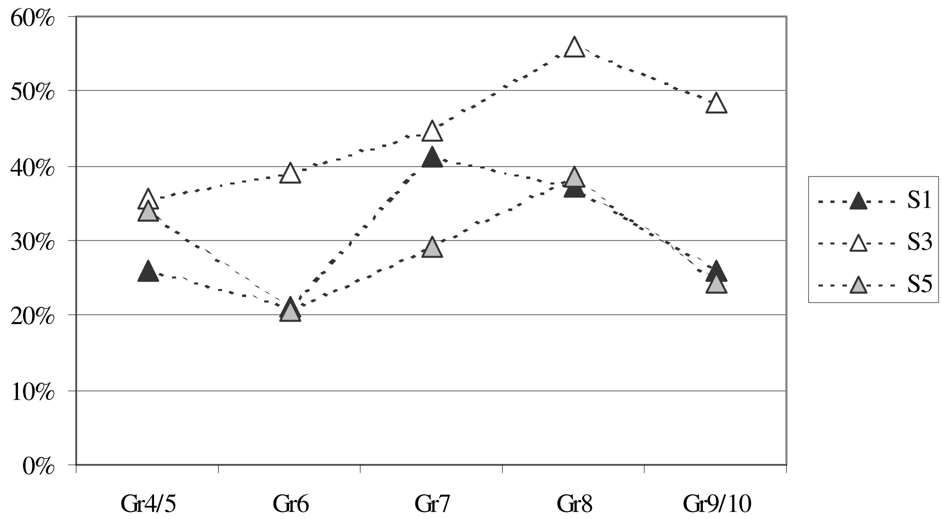


Figure 6.2: Persistence in S over one semester from S1, S3 and S5 by grade of current test

To further explore the increased persistence in S in the secondary school, Figure 6.3 provides the persistence by school group; grouping the grades to maintain reasonable sample sizes. Clearly the persistence in S in Grades 7 and 8 is higher than the other grades and this is not confined to just one school group. Therefore, this phenomenon is not due to particular teaching by a few teachers, it is more likely that the major factor is the curriculum; in particular we propose the various reasons related to exposure to negative numbers and negative powers as explained in section 2.2.

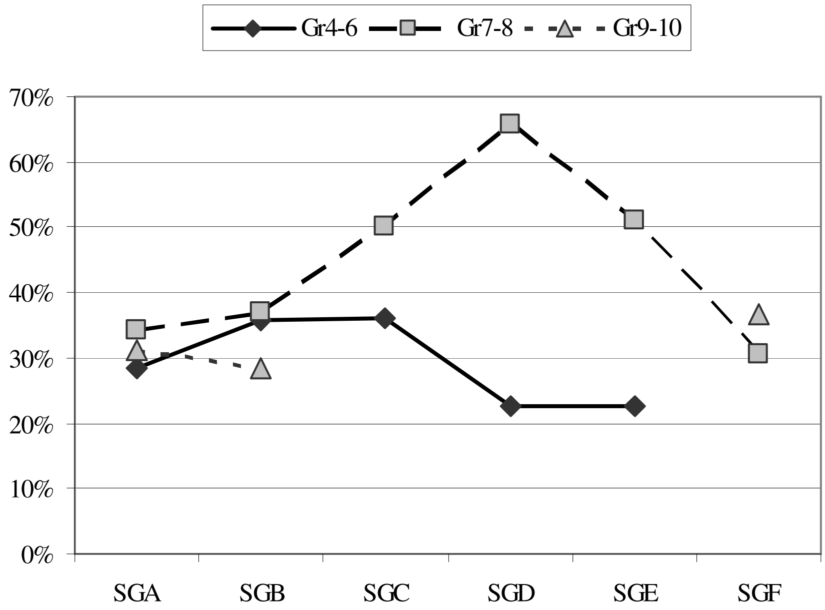


Figure 6.3: Persistence in S over one semester by school group, by grade of current test

6.4.2 Variations with school group

The following discussion considers variations by the school groups in the persistence in L, S and U; as SGF has only secondary students, the most appropriate analysis also uses school level. Figure 6.4 indicates that the higher persistence in L in primary schools (compared with secondary schools) is not due to one school group in isolation. Likewise, the higher persistence in S in secondary schools (compared with primary) is not due to one school group in isolation. In particular, several school groups have very high persistence in L (over 50% in primary schools in SGA and SGC), and several school groups have very high persistence in S (50% or more in secondary schools in SGC, SGD and SGE).

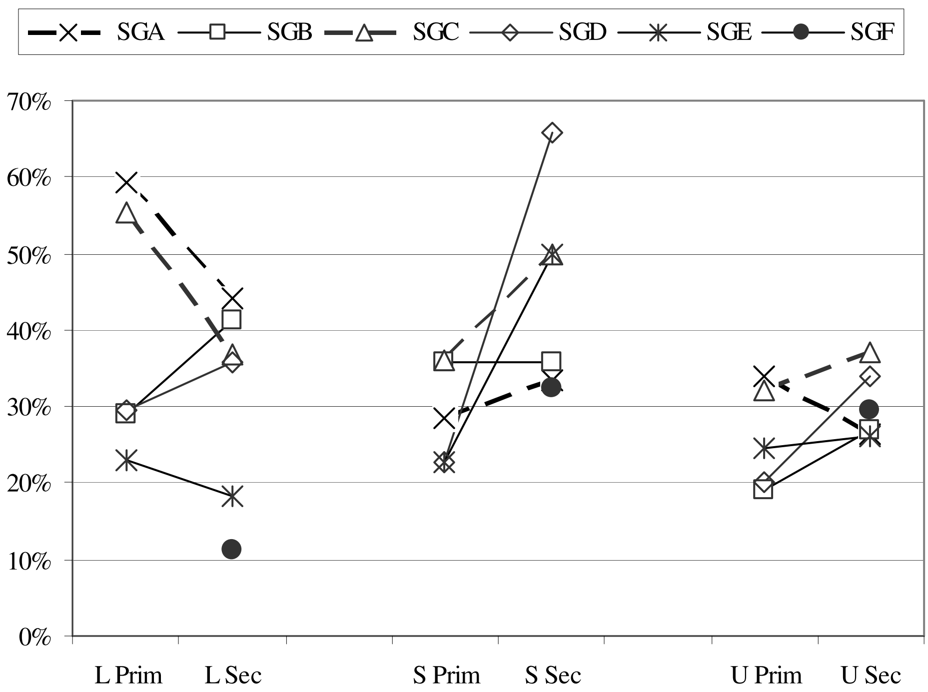


Figure 6.4: Persistence in L, S and U for primary and secondary samples by school group

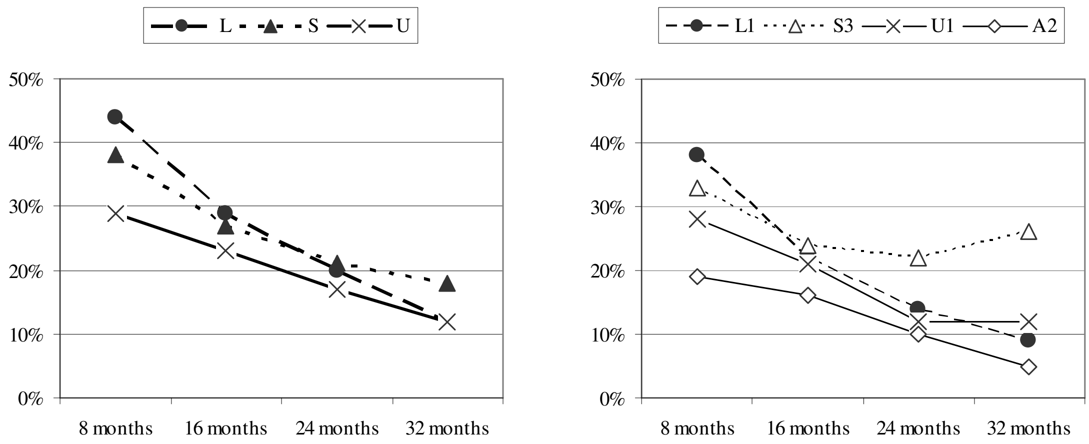
6.5 Persistence over various semesters

The persistence discussed above was calculated using pairs of consecutive tests; these transitions will be labelled as T_iT_{i+1} to indicate that the calculation was based on combining data from consecutive tests T_1T_2 and T_2T_3, .., T_6T_7. Following students from their 1st test to their 3rd test provides information on pairs of tests by a student approximately 16 months apart. Hence, T_iT_{i+2} will indicate that the calculation was based on combining data such as this: T_1T_3 and T_2T_4, .., T_5T_7. Similarly, we can define T_iT_{i+3} and T_iT_{i+4}. Hence, these measures are related to later occurrences of the codes, which may or may not be in sequences.

Figure 6.5a provides the percentage of students retesting in the same coarse code (L, S or U) after various periods of time (1 to 4 testing periods), and Figure 6.5b provides similar information for selected fine codes.

Figure 6.5a indicates that after three testing periods (a period averaging 24 months or 2 years), approximately 20% of the students who tested as L have additional occurrences of L; similarly for students with S and U. (Stacey and Steinle, 1999b, had this result on a slightly smaller sample).

Figure 6.5b indicates that after a period of two years (3 testing periods), approximately 15% of L1 students have additional L1 occurrences, about 10% for U1 and A2 students and over 20% for S3. The last data point for S3 in Figure 6.5b suggests that more students return to S3 after a longer period of 32 months. As the sample size for this calculation is reasonable (n=43 students) this result cannot be ignored. S3, the code which makes errors with equal length decimals, is persistent.



a) L, S and U b) L1, S3, U1 and A2

Figure 6.5: Persistence in various codes after various semesters

It should be noted that another analysis confirmed these results. For the 972 students who tested as L at some stage in their test history, the date of their first L test was recorded and tests completed more than two years after this date were examined. Of the 362 students were still involved in the project, 74 completed an L test after the two year period. Hence, 20% of those still involved in the project two years after their first L test were still completing tests allocated the code L. This analysis was repeated with S students, and 29 of the 178 students (16%) still involved in the project were completing tests coded as S.

6.6 Conclusion

Clear evidence has been provided of students who retain the same misconception from one test to the next, as they move from one grade to another throughout their time in school. In particular, there were twelve students who completed a sequence of five tests allocated the same code; their test histories contained [L1,L1,L1,L1,L1] or [S3,S3,S3,S3,S3] demonstrating unchanged decimal misconceptions over about three years.

While it might be predicted that L misconceptions are more a primary school phenomenon, a considerable number of students were testing as L into secondary school. This is tempered by the decreasing tendency for students to persist in L on their next test; hence these misconceptions are not very likely to be found in adults.

It has been demonstrated, however, that S misconceptions are more a secondary school phenomenon. The increased tendency for students in Grades 7 and 8 to persist in S3 in the next semester suggests the influence of teaching; most likely confusion with negative numbers or scientific notation. It seems reasonable that some students will leave secondary school with one of the S misconceptions; recall that both Grossman (1983) and Putt (1995) found about 40% of their tertiary samples chose a distractor which corresponds to S behaviour. The increasing tendency (with grade) for students to persist in A2 in the next semester suggests that these students are becoming more comfortable with their understanding; possibly constantly rounding the results of calculations to two decimal places reinforces their belief that the decimal numbers form a discrete system. The code A2 will be investigated further in later chapters, in particular, Chapter 8 Regression.

CHAPTER 7 HIERARCHY

7.1 Introduction and terminology

This chapter discusses the issue of hierarchy. Is it possible, from analysing this dataset, to order the codes from the most *primitive* to the most *advanced*? Does this order confirm or contradict the links between the codes and the ways of thinking they are purported to represent? It is proposed that any such ordering of the ways of thinking does *not* represent a series of steps or stages that a student must pass through on the way to expertise.

As noted in Chapter 3, the Decimal Comparison Test is a diagnostic test and, unlike "typical" tests, the percentage of correct answers by a student is a meaningless quantity. It is not sensible to accord students a position in a hierarchy based on the percentage of correct answers. To explain this, we need to consider a typical test.

It is usually possible, in a typical test, to rank the items from the "easiest" item (with the highest facility) to the "hardest" item (with the lowest facility). An item that is easy for younger students will be even easier for older students. Changing the mix of items within a test might make the test easier or harder, but the relative positions of two students (as determined by their relative scores) would remain the same.

This diagnostic test does not fit into this expected scenario. Firstly, there are items that are easy for younger students but harder for older students! The Type 4 items (such as 4.4502 / 4.45) are very easy for students with whole number thinking (who readily choose 4502 as larger than 45), but are much harder for older students where a higher proportion of non-expert students complete tests which are coded S or A2 and therefore make errors on this item type.

Secondly, it is possible to construct two different versions of the Decimal Comparison Test, with a different mix of items, in which the relative positions of two students change. On the current test with 30 items, a typical student coded as L1 scores 19 and a typical student coded as S3 scores 13. By increasing the proportion of Type 1 items (compared with Type 2 items), it is possible to create a test that a student exhibiting S behaviour scores highly, (and hence a student exhibiting L behaviour scores poorly). If the proportions are reversed, the student exhibiting S behaviour will score more than the student exhibiting L behaviour. Hence, the relative position of students, based on their percentage of correct answers, is a function of the mix of items on the test, and is therefore of no use when attempting to determine a hierarchy.

Terminology

By considering the typical ordering of two codes within students' test histories, we can create a *developmental ordering* of the codes. For example, in almost every case that a student has both L1 and A1 codes in their test history, the L1 code occurred before the A1 code. Hence, L1 is *developmentally younger* than A1, or A1 is *developmentally older* than L1. All pairs of codes will be considered to attempt to produce a developmental ordering of codes.

The *proximity to expertise* is the conditional probability of moving to expertise on the next text, given that the student changes from their current code. It was found in Chapter 6 that some codes were hard for students to leave; such students appear to have been "side-tracked" and seem to be reasonably comfortable with their current way of thinking. Due to the higher persistence of students in some codes, students who are "stuck" in a code will not be included in these calculations. In other words, the rate to expertise in this chapter will be conditional on the student changing codes (i.e. not persisting in their current way of thinking). This procedure will provide a ranking of the codes that indicate their relative proximity to expertise.

In section 7.2, previous research that investigated or commented on the issue of hierarchy will be discussed. Section 7.3 contains the derivation of the developmental ordering of the codes (both coarse and fine). In section 7.4, the codes will be ranked according to their proximity to expertise; students who are more likely to move to expertise on the next test are therefore considered to be more advanced than students in other codes who are less likely to move to expertise on their next test. The ranking of the codes based on the developmental ordering will be compared with that from the proximity to expertise in an attempt to determine a hierarchy.

7.2 Previous research

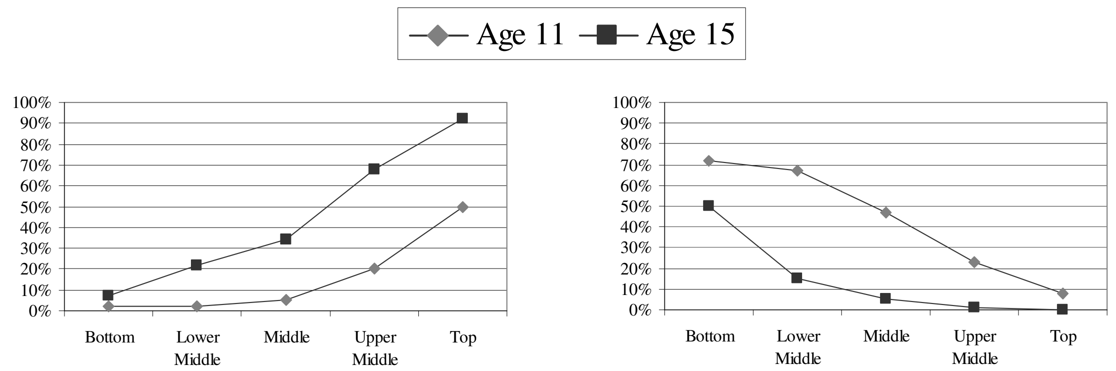
Some researchers have proposed that particular misconceptions are more advanced or more primitive than others by a theoretical consideration of the knowledge upon which the misconceptions are based. Other researchers have used empirical data from additional tasks given to students; the analysis of the facilities by students allocated to various misconceptions provides another basis for determining a hierarchy. As in Chapter 2, the terminology of Resnick *et al*., (1989) (i.e. Whole Number Rule, Zero Rule, Fraction Rule and Expert Rule) will be used to describe results from a group of earlier researchers.

Sackur-Grisvard & Leonard (1985) and Resnick *et al*., (1989) made predictions (based on their statements about the causes of the misconceptions) that the most primitive rule was Whole Number Rule; the Fraction Rule was more advanced and the Expert Rule (obviously) was considered to be the most advanced. The position of the Zero Rule above the Whole Number Rule in a hierarchy followed from the definition of the Zero Rule (i.e. that it is based on the Whole Number Rule, with an additional piece of knowledge). Sackur-Grisvard & Leonard proposed that the Zero Rule was more advanced than the Fraction Rule, although Resnick *et al*. (based on interview data with a small sample of Grade 5 and 6 students) concluded that Fraction Rule is more advanced than the Zero Rule.

Moloney (1994) used additional written tasks with students from Grades 4 to 10 who were allocated to these various rules. The majority of the Whole Number Rule students (over 80%) were in Grades 4-6; hence a comparison of the performance by rule use is affected by age distribution. The general trend within the facilities on the items, (see Table 2.14), however, suggests this same hierarchy (from lowest to highest): Whole Number Rule, Zero Rule, Fraction Rule and then Expert Rule.

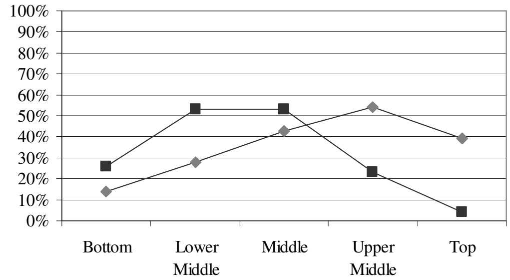
Moloney and Stacey (1997) conducted the first longitudinal study and found very little movement of students between classifications after one year of schooling. Four of the 12 students who were Unclassified moved to Expert Rule, compared with only two of the 15 Fraction Rule students; this ranking of relative likelihood of moving to expertise has been confirmed in Chapter 5. None of the 6 Whole Number Rule students became experts. Although based on small numbers, the relative movement to expertise suggests this same hierarchy with the extra information that students who are unclassified are the most likely to move to expertise.

Foxman *et al.* (1985, p56) suggest that *largest-is-smallest* error (roughly equivalent to S behaviour) is more advanced than *decimal-point-ignored* error (roughly equivalent to L behaviour). Based on the overall performance on a wide variety of test items, students were allocated to one of five attainment bands. The analysis of the percentage of responses in each attainment bands at the two different age groups provides a rather complex picture of the S behaviour. To illustrate this complexity, the percentages from Foxman *et al.*, (page 57) have been used to create Figure 7.1.



a) Correct answers

b) L behaviour



c) S behaviour

Figure 7.1: Prevalence of Correct, L and S behaviour in APU attainment bands for two age groups, from Foxman *et al.* (1985), p57

The distributions of the correct and L behaviour (Figures 7.1a and b) follow expected trends. Firstly, within each age group, there are *increasing* proportions of correct responses (corresponding to *decreasing* proportions of L behaviour) across the bands. Secondly, within each band, there are *increasing* proportions of correct responses (corresponding to *decreasing* proportions of L behaviour) across the age groups.

There are unexpected trends associated with the S behaviour, however, that can be seen in Figure 7.1c. Firstly, within each age group, there are *increasing* proportions of S behaviour from the Bottom band to the Middle band (and for 11 year olds, to the Upper Middle band). Secondly, within the Bottom band to the Middle band, there are *increasing* proportions of S behaviour from the older students.

These results show quite clearly that the S behaviour is unlike the L behaviour and is not tending to disappear with age.

That S behaviour is more advanced than L behaviour is supported by changes after instruction; Fuglestad (1998) found that the increase in expertise was accompanied by a decrease in the prevalence of L behaviour and a slight increase in the prevalence of S behaviour. Note that as the longitudinal results were not provided, we cannot determine whether the same students retained their S behaviour, (with a few new students joining in) or whether the original students with S behaviour moved elsewhere, and were replaced by all new students.

Results of preliminary analyses, interpreting hierarchy as the relative movement of coarse codes to A on consecutive tests, within this data, have been published in Stacey and Steinle (1999a and 1999b), and the corresponding results for the complete dataset appeared in Chapter 5 (Table 5.16). In general, about one quarter of the L students move to A on their next test, compared with about one third of the S students and one half the U students. Based on readiness to move to A, these results confirm the hierarchy suggested by Moloney and Stacey's (1997) results, that is, lowest to highest: L, S, U then A. Note that as U indicates the complement of the other classifications, it could be argued that it does not deserve a mention in this chapter. It will be included, however, for completeness.

The following analyses in this chapter will determine whether this hierarchy (L, S, U then A) holds for students of all grades and all school groups and then determine the hierarchy within the fine codes. Due to the unexpected results in the Chapter 6, (i.e. the increase in persistence of secondary students in A2, S3 and S5, see Table 6.11) this measure will be calculated on the reduced sample, thereby answering the question: *Given that a student moves from their current code, what are the rates to expertise?*

7.3 Developmental ordering of the codes

In this analysis, the order that the codes appear in the test history of each student will be considered. Table 5.13 provided the number of students with any two given coarse codes (not necessarily consecutive) but without any indication of which code came before the other; Table 5.14 provided the same information for the fine codes. This analysis will now consider which codes came earlier and which came later within students' test histories. As this procedure is based on full test histories it is therefore not possible to provide any results by grade; this is possible, however, with the alternative measure which will be presented in section 7.4.

A decision needs to be made regarding students with several occurrences of a code. For example, if a student has a test history of [S,U,L,U], which U will be considered? Will we conclude that the U comes before or after the L? The decision was made to choose the first occurrence of a code. Note that if later occurrences of a code are in a sequence, then this analysis is not compromised. Scattered occurrences, however, will result in tests being ignored. (For U1, 9% of involved students have scattered occurrences, while the remaining 10 fine codes have an average of 4%, see Table 6.9.) So a student with a test history of [S,U,L,U] would contribute three times to this analysis as follows:

- the first occurrence of S precedes the first occurrence of U
- the first occurrence of U precedes the first occurrence of L
- the first occurrence of S precedes the first occurrence of L

While the first two points would have been counted in earlier analyses of consecutive test pairs, the third point would not have been. For the remainder of this section, the term "occurrence" will refer to a student's first occurrence of that code.

7.3.1 Coarse codes

This procedure, considering the order of the first occurrence of two different coarse codes, was used on every student's test history, and the results are provided in Table 7.1. For example, the student discussed above would contribute three times and be included with: the 161 students with S before U, the 57 students with S before L, and the 57 students with U before L.

Table 7.1: Number of students with any 2 coarse codes, row before column

Occurs before	Occurs after			
	A	L	S	U
A		33	81	177
L	462		160	254
S	342	57		161
U	348	57	97	

Consider the pair of codes, A and L. There are 495 students with both L and A *in any order*, (see Table 5.13) and Table 7.1 indicates that 462 students (93%) have L before A, and the remaining 33 students (7%) have A before L. This calculation was repeated for the other five pairs of coarse codes and the results provided in Table 7.2.

Note that, by definition, there are no entries on the main diagonal of Table 7.2, (although if there were, they would be 50%) and the sum of the entries corresponding to the ordering XY and YX is 100%. The rows and columns have been ordered so that the entries increase down a column, and the entries decrease across a row. As a consequence, all the entries below the main diagonal are greater than 50%.

Table 7.2: Percentage of students with any 2 coarse codes, row before column

Occurs earlier	Occurs later			
	A	U	S	L
A	*	34	19	7
U	66	*	38	18
S	81	62	*	26
L	93	82	74	*

The resulting ordering of the rows and columns (L, S, U then A) confirm that A is the "oldest" code as all the other codes are at least twice as likely to occur *before* A as *after* A, (entries of 66% or more in the first column). Note that this procedure did not make any assumptions *a priori* regarding which code should be the oldest code; this was determined by this analysis.

Similarly, the entries in the last row indicate that L is the "youngest" code as all the other codes are at least twice as likely to occur *after* L as *before* L. The code S occurs before U more often than U before S (62% compared with 38%), although the ratio here is only 1.6 (rather than 2 or more as noted above). Hence, the developmental ranking of the coarse codes is L, S, U then A which confirms the ordering obtained by considering the relative rate to expertise in Chapter 5 (Table 5.16).

7.3.2 Fine codes

This procedure was then applied to every student's test history using the fine codes. For example, of the 283 students with both A1 and L1 (see Table 5.14) only 5 students had their first A1 before L1, while the other 278 had their first L1 before their first A1. Table 7.3 contains the number of students with each possible pair of fine codes, row before column.

Table 7.3: Numbers of students with any 2 fine codes, row before column

Occurs	Occurs after										
Before	A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
A1		92	59	5	19	3	17	23	23	107	28
A2	121		23	3	7	2	9	17	8	28	8
A3	90	18		5	6	3	5	4	4	35	7
L1	278	35	63		83	66	55	54	32	191	3
L2	106	16	17	14		5	19	15	10	49	3
L4	94	18	19	21	24		12	14	13	41	3
S1	117	30	20	7	8	4		27	23	36	3
S3	124	23	25	18	14	10	28		26	70	10
S5	96	25	16	13	6	5	26	33		62	10
U1	302	41	54	28	37	24	43	50	31		14
U2	19	2	4	1	0	0	1	2	1	9	

As before with the coarse codes, percentages are now created and provided in Table 7.4. Asterisks denote the main diagonal, for which there are no entries, and only percentages based on samples of 16 or more are shown in this table; the 6 cells with samples between 16 and 20 are indicated by entries which are in bold and italics. Unlike the situation with the coarse codes, there is no clear ordering of the codes to produce the required trends in the columns and rows. A useful compromise has been to order the fine codes according to the ordering of the coarse codes (L, S, U then A) and then an attempt has been made to place the percentages over 50% (within the shaded cells) under the main diagonal.

Table 7.4: Percentages of students with any 2 fine codes, row before column[1]

Occurs	Occurs later										
before	A1	A3	A2	U2	U1	S1	S3	S5	L2	L4	L1
A1	*	40	43	60	26	13	16	19	15	3	2
A3	60	*	44		39	20	14	20	26	14	7
A2	57	56	*		41	23	43	24	30	10	8
U2	40			*	39						
U1	74	61	59	61	*	54	42	33	43	37	13
S1	87	80	77		46	*	49	47	30	***25***	11
S3	84	86	58		58	51	*	45	48	42	25
S5	81	80	76		67	53	55	*	***38***	***28***	28
L2	85	74	70		57	70	52	***63***	*	17	14
L4	97	86	90		63	***75***	58	***72***	83	*	24
L1	98	93	92		87	89	75	72	86	76	*

1: not shown if n<16, bold & italic for n<20, (6 cells)

Firstly consider the shaded cells in Table 7.4 which correspond to the fine L codes. All 6 entries associated with these three codes indicate that there is a clear developmental trend from L1 to L4 to L2. For example, L1 is more than six times as likely to occur earlier than, rather than later than, L2. Similarly, L1 is more than three times as likely to occur earlier than, rather than later than, L4. Furthermore, L4 is more than four times as likely to occur earlier than, rather than later than, L2.

Now consider the shaded cells in Table 7.4 which correspond to the three fine A codes. All 6 entries associated with A1, A2 and A3 are between 40% and 60%, hence, while A1 is "older" than both A2 and A3, the differences are small. In fact, it will be shown in the next chapter that the reason for this is the considerable number of students who move from A1 to either A2 or A3 (an example of *regression*). Hence, the fine A codes will be investigated more fully in the next chapter.

Now consider the shaded cells in Table 7.4 which correspond to the three fine S codes. There are no clear developmental trends; all 6 entries associated with S1, S3 and S5 are between 45% and 55% and hence no conclusions can be made about which codes are younger or older.

The cells of Table 7.4 which correspond to fine L and fine A indicate that all the L codes occur earlier than all the A codes. A similar result holds for all the S codes which occur earlier than all the A codes. There is one value of 58% (S3 before A2) which is rather small compared with the surrounding entries, (eight of these cells have entries of over 75%). Hence, while S is an earlier code than A, more students than expected have their first occurrence of S3 *after* their first occurrence of A2. Two possibilities are offered to explain this observation. Firstly, it will be shown in Chapter 8 that A2 students are likely to be *harbouring* S behaviour, which is temporarily suspended by the use of an algorithm for comparing decimals. Secondly, S3 may be a combination of two ways of thinking which occur at different times, one younger and one older; this is discussed in Chapter 9.

The code S3 is involved in other lower than expected entries in the cells which correspond to fine L before fine S; only 52% of L2 tests occur before S3 and only 58% of L4 tests occur before S3. Hence, the code S3 needs further investigation.

Now consider the codes U1 and U2. There are 47 students with both A1 and U2 and 60% of these students have A1 before U2 in their test history; hence U2 appears to be an older code than A1. The small number of U2 codes allocated means that there are not sufficient numbers of students with various combinations of U2 with other codes to investigate other aspects of the hierarchy, and hence, this code is investigated more thoroughly in the next chapter.

Hence, this first measure of hierarchy produces an ordering of the fine codes (lowest to highest) of L1, L4, L2, S5, S3, S1, A2, A3, A1, although the fine S codes are very close. The next section will consider another measure of hierarchy that can be imposed on the dataset and these two different measures will then be compared in section 7.5.

7.4 Proximity to expertise

Comparison of the relative rate to expertise on the next test (see Table 5.16) suggests a hierarchy of L, S, U then A. As this rate is based on pairs of tests it can therefore be calculated for students at a particular grade. As discussed above, it requires a code to be defined as expertise. For the fine codes, this is obviously A1; for convenience, the coarse code A is used in the preliminarily analysis in 7.4.1 to indicate expertise.

To remove the complication that some codes have high persistence, only students who are changing their code are involved in the calculation. We might expect that, in general, as students get older, they have an increased tendency to move to expertise. This is tempered, however, by the creation of a smaller sample of older students who have not achieved expertise and who may often be referred to as "slow learners". We might also expect that there are particular grades that have increased rates of movement to expertise, due to teaching that is focussed on decimal notation. Another alternative is that new teaching can interfere with current understandings as mentioned in Chapter 2, and this will also be investigated.

7.4.1 Coarse codes

As mentioned in section 7.2, the relative rates to A on the next test were determined in Chapter 5 and the first question is: *Does the ranking of L, S, U then A, determined from the rates to expertise, hold for students of all ages and from all school groups*? Figure 7.2 provides the proximity to A from an L, S and U test completed in either primary school or secondary school. (Note that these new figures are higher than those in Table 5.16 due to the restriction to students who change code.)

While the students in secondary school who completed either U or L tests were closer to A, than their counterparts in primary school, this is not true of the S students. Again we find that students exhibiting S behaviour do not fit predictions and will be investigated more thoroughly below.

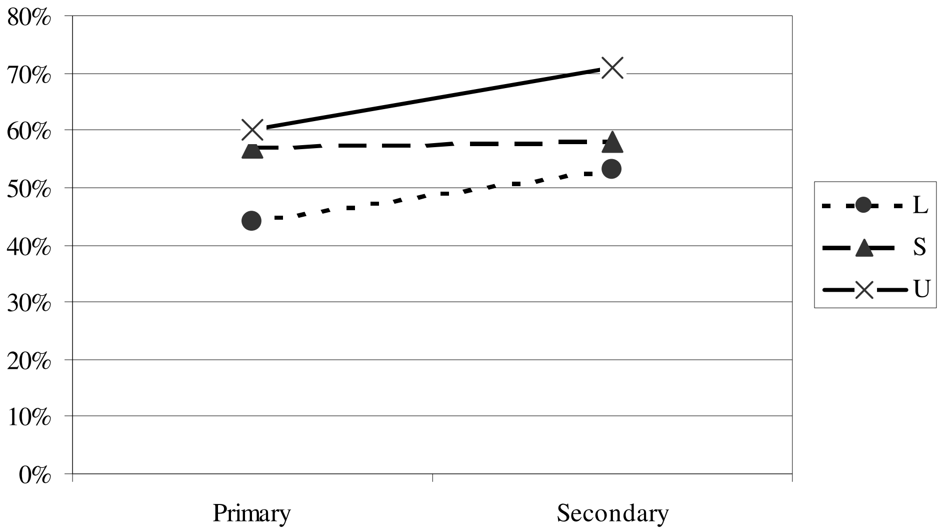
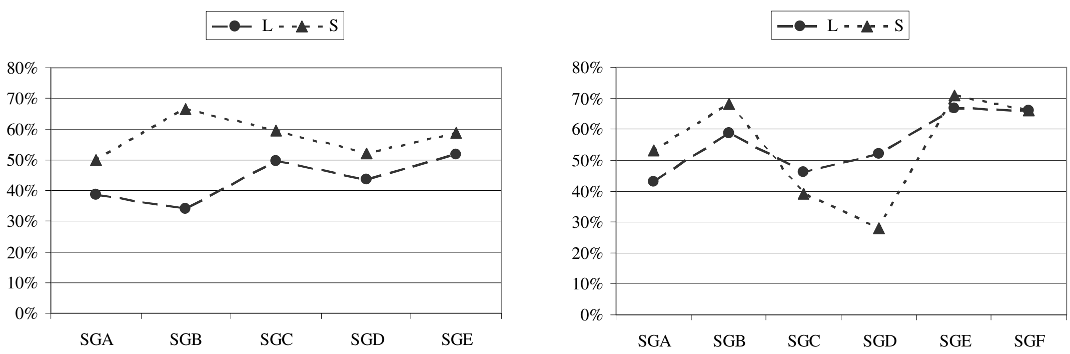


Figure 7.2: Proximity to A from an L, S and U test, by primary and secondary level

The following investigation into the proximity to A will focus on L and S as the position of U was found to be consistently above both L and S in these analyses.

Figures 7.3a and 7.3b demonstrate the variations in the proximity to A by the various school groups in the two school levels. It is clear that the proximity to A from L is less than S in the primary schools, but some markedly low figures from S occur in several of the secondary schools. Hence, the position of S above L in this ranking of readiness to move to expertise does not hold for all secondary students and a more detailed investigation (using fine codes) is required.



a) Primary b) Secondary

Figure 7.3: Proximity to A from L and S, by school group and school level

7.4.2 Fine codes

In this section, expertise is defined to be A1. The proximity to A1 was determined for each of the fine codes and the resulting ranking of the fine codes is provided in Table 7.5. In general, the fine codes are ranked as per the coarse codes, that is, L, S, U then A. The one exception is that L2 is interspersed within the S codes.

The next question to be answered is: *Does this ranking hold for all students?* Figure 7.4 illustrates the above ranking of the fine codes by the proximity to A1 for the primary and secondary samples. (There is no point for primary U2 tests due to small numbers.) Some fine codes have very similar values for proximity to expertise in the primary and secondary samples. Some other codes have a higher proximity to expertise in the secondary sample. What is unexpected, however, is that for the codes A2, S1 and S3, there is a considerable *decrease* in the proximity to expertise from primary to secondary. While it was determined in the previous chapter that both A2 and S3 have higher rates of persistence in secondary compared with primary, this is not the reason for the drop in Figure 7.4, as all rates within this chapter are based on students who change codes. This is therefore a different phenomenon.

Table 7.5: Ranking of fine codes based on proximity to A1

Fine Code	Number of suitable consecutive test pairs	Proximity (%) to A1*
A2	n=227	65
A3	n=164	57
U2	n=39	56
U1	n=548	51
S1	n=213	38
L2	n=201	36
S3	n=258	33
S5	n=202	31
L4	n=167	30
L1	n=531	24

**rate to A1 given change of code*

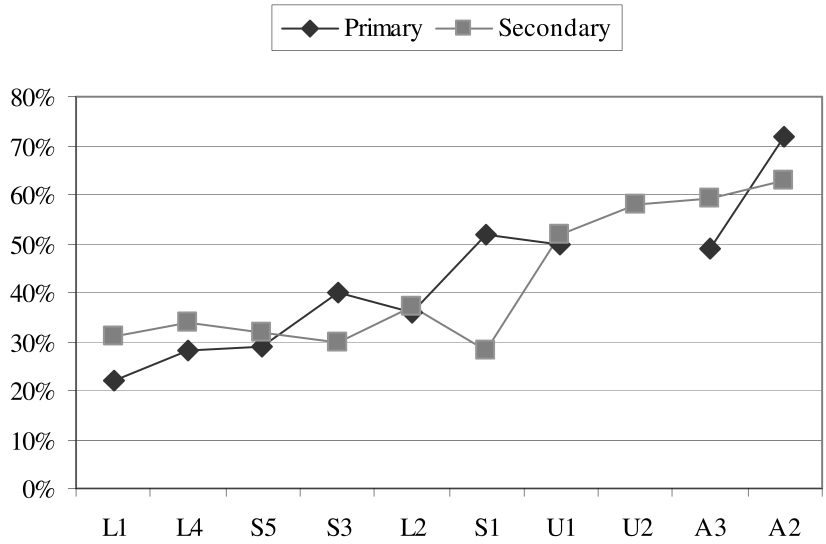


Figure 7.4: Proximity to A1 for the fine codes for primary and secondary samples

Ranking the fine codes for the primary sample and then for the secondary sample provides different orders than appear in Figure 7.4 due to the drop by S1 and S3. (U2 is not provided, as there were few U2 tests in primary school). These rankings are provided in Table 7.6.

Table 7.6: Ranking of fine codes based on proximity to A1 by school level

Fine Code	Primary Proximity (%) to A1		Fine Code	Secondary Proximity (%) to A1	
A2	72	(n=53)	A2	63	(n=174)
S1	52	(n=91)	A3	59	(n=127)
U1	50	(n=234)	U1	52	(n=314)
A3	49	(n=37)	L2	37	(n=95)
S3	40	(n=96)	L4	34	(n=62)
L2	36	(n=106)	S5	32	(n=130)
S5	29	(n=72)	L1	31	(n=147)
L4	28	(n=105)	S3	30	(n=162)
L1	22	(n=384)	S1	28	(n=122)

Consider the relative position of the three L codes in the primary and secondary lists above. The order L1, L4 then L2 appears in both lists; this measure thus reinforces the developmental ordering established earlier.

The situation within the S codes is more complicated. In the primary sample, S1 is one of the most advanced codes, while it drops to the very bottom of the list in the secondary sample, even below L1. Hence, within the primary sample, the ordering of the fine S codes is (lowest to highest) S5, S3, S1, in contrast to the ordering within the secondary sample of S1, S3, and S5. These changes are not due to differences in S5, but rather, to large drops in both S1 and S3.

To further investigate the three codes (A2, S1 and S3) which have lower proximity to A1 for the secondary sample (compared with the primary), Figure 7.5 provides the proximity to A1 by the grade of the current test. (Table 1 in Appendix 9 contains the sample sizes for these points.)

The code A2 has a reasonably constant proximity to A1 of about 70% with the exception of students in Grades 9 and 10. The 40 students involved in this older sample have a lower proximity to A1 of only 45%. Inspection of Tables 12 and 14 in Appendix 7 indicates that the A2 students who do not persist and do not move to A1, tend to move to A3 or U1, and the A3 students who do not persist and do not move to A1, tend to move to U1.

The code S1 has more erratic variations with a high in Grade 6 (nearly 60%) dropping to half this in Grades 7 and 8. Code S3 has two rates, about 40% for Grades 4 to 7 and then about half this for Grades 8 to 10. Hence the lower rates from S to A by students in secondary school (noted earlier this chapter) are due to lower rates to A1 by both S1 and S3. Such variations provide weak evidence that the codes S1 and S3 are both being allocated to students with various different ways of thinking (discussed further in Chapter 9).

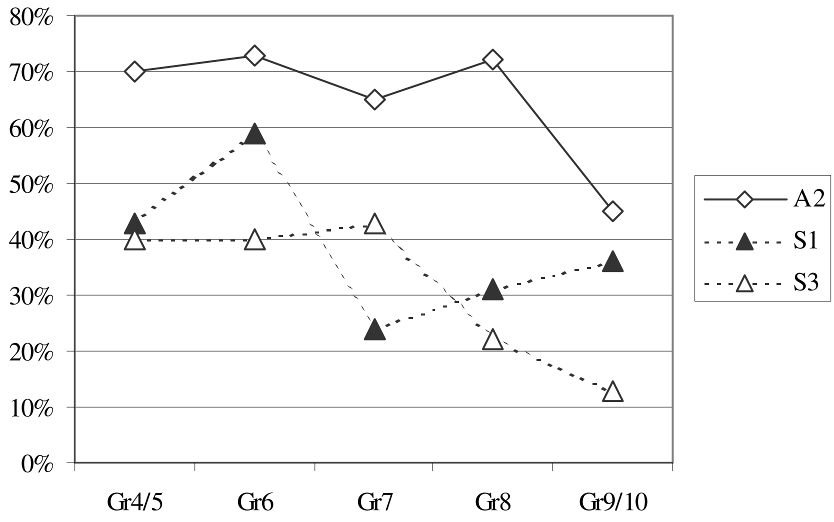


Figure 7.5: Proximity to A1 for A2, S1 and S3 by grade of current test

7.5 Conclusion

Two measures were introduced and used in this chapter to explore the issue of hierarchy. The developmental trends were calculated by considering (for students with two given codes) which code tended to occur earlier in students' test histories. The proximity to expertise was determined by the relative rate to expertise on the next test, given that the student did not remain in the same code. Both these measures gave the same ordering for the coarse codes; that is, L, S, U then A. Whilst this ordering seems like a clear result, examination of students in secondary schools indicates that the proximity of S to A decreases to such an extent that S replaces L at the lowest position. (Note that if this result was confined to a particular school group, then the explanation might be particular teaching in one school, but this is not the case.)

Both these measures were used on the fine codes and were in agreement that, within the L codes, L1 is consistently the lowest; L2 consistently the highest; and then L4 is between these two codes. That L2 (*column overflow thinking* or *zero-makes-small thinking*) is more advanced than L1 (*whole number thinking*) is predicted from a theoretical consideration of the ways of thinking in Chapter 3. In fact, had the test distinguished the two ways of thinking in L2 we would have predicted differences between them too.

Neither of the two measures provided a clear result for the ordering of the fine S codes. The analysis of development trends found that S1, S3 and S5 were all equally likely to occur earlier or later in students' test histories, and the proximity to expertise for S1 and S3 for secondary students dropped to such an extent to be further from A1 than L1 (previously assumed to be the most primitive code). To explain these unexpected decreases it is proposed that each of the codes S1 and S3 result from two different ways of thinking as discussed in Chapter 3 which vary in proportions between primary and secondary school students.

Similarly, the relative positions of A2 and A3 in a hierarchy are not clear. While the proximity to A1 indicated that A2 was above A3 in the hierarchy, the developmental ordering found that the first occurrence of A2 tended to be before the first occurrence of A3 more often than not (56% A2 first, and 44% A3 first). It is proposed that it is possible that a student with a deficient understanding of decimal notation, (e.g. two whole numbers separated by a decimal point, with the largest number on the right being 99) are completing tests coded as A2, yet this is not particularly sophisticated thinking. It is proposed that it is "sophisticated enough" to get by in many situations. More will be learnt about A2 and A3 students in the next chapter, Regression.

The adequacy of these measures to determine a hierarchy is now discussed. A theoretical consideration of the ways of thinking behind the codes, predicts that L2 is more advanced than L1. Both measures used in this chapter provide this same ordering of L2 above L1, and hence, the measures are validated. The developmental trends identify the codes that occur earlier and later for each student, and these have not been used in isolation to determine a hierarchy, as we believe that older does not necessarily mean "cleverer".

Recall that the code U does not refer to a misconception; rather it is a collection of non-expert tests that do not fit into the predicted misconceptions. It is not, therefore, a group of students with one particular way of thinking. Its place above L and S, in the relative rates to A, provides us with the confirmation that, of the students making errors, those who are *not* holding onto a misconception are those who are most likely to move to expertise on their next test. Consider a student who uses a mixture of strategies as they complete the test, and is likely therefore to be coded as U. As they complete the DCT, they may notice that they are using a mixture of strategies and that they are unsure of which strategy to use on any one item. For example, recall that L students get a score of 0 or 1 correct on the 5 Type 1 items. Steinle and Stacey (2003a) found that the two items in Type 1 which L students are most likely to choose correctly had non-zero integer part. Hence, for some students this feature generated a different strategy to the usual choice of the longest decimal is the largest. A student who answered only these two items correctly in Type 1 would have a score of 2 and hence their test would automatically be coded as U.

So, a student completing a test coded as U may be using a mixture of strategies and as they complete the 30-item test, they may come to the realisation that there is something that they need to learn. This may be a crucial moment for such a student and could result in increased attention in subsequent classes; hence an increased likelihood of testing as an expert on the subsequent test.

CHAPTER 8 REGRESSION FROM EXPERTISE

8.1 Introduction and terminology

Another noteworthy phenomenon observed in data analysis is the case of students able to answer the Decimal Comparison Test as an expert on one occasion but not on the following test; this phenomenon is referred to as *regression*. Note that cross-sectional studies will not detect this phenomenon; it requires a longitudinal study such as this.

As found in Chapter 5, the overall rate of re-testing as an expert was close to 90%, so that 9 out of 10 students retain their expertise. On the other hand, 1 in 10 students regressed from expertise on their next test. The effects of age and school group on this rate will be considered. Do younger and older students have the same rate of regression? What more can we learn about students who complete tests which are coded as expert? In particular, can we learn anything that might distinguish those with good understanding from those who follow a rule that is not understood? Are the latter the 10% who regress?

Three scenarios will be proposed that may cause regression. Firstly, it would be very easy for a teacher, or another student, to assist a student complete this test as an expert, merely by reminding them of one of the algorithms (discussed in detail this chapter) that can be used to compare decimals. Although we have no evidence that this occurred, we are aware that feedback provided to teachers was used, at the primary school in SGD, as “league tables”. Does this school have higher rates of regression than other schools? Our hypothesis is that after students forget a particular algorithm, they will return to the same non-expert state that they were in previously.

The second situation that might result in regression is the use of inadequate models to explain decimal notation. As discussed in Chapter 2, some models (such as dollars and cents) provide such a limited context that students are unable to correctly (and consistently) compare some pairs of decimals (e.g. they may think 4.45 and 4.4502 are equal). In particular, models which are limited to two decimal places may be leading to two of the incomplete algorithms described later (rounding/truncating to two decimal places).

Thirdly, as also noted in section 2.2, new teaching can interfere with a student's performance on tasks that had previously been completed correctly. Consider a student who has completed a test coded as A1. Later, when other new knowledge is being learnt, this new knowledge can interfere with existing knowledge. Specifically, when a student is introduced to negative numbers or negative powers of ten in Grade 7, they need to integrate these new numbers into their current understandings of numbers. In particular, noting that -6 is greater than -75, this may encourage a student to consider 0.6 to be greater than 0.75, i.e. to treat the decimal portion as a whole number and then choose the smaller whole number as the larger decimal. In this case we predict that a confused student may lose expertise and move to S behaviour.

Terminology

For this chapter, *expertise* will be defined using the fine classification (A1) rather than the coarse classification (A) as has been done in some of the analyses in earlier chapters. A student is said to *regress* if they have a test coded as A1 followed by a test with a different code. Hence, regression is detected when two consecutive tests by a student are examined. The *regression pair of tests* refers to these two consecutive tests, that is, an A1 test followed by a non-A1 test. The first test in this pair is then referred to as the *unstable A1* and the second test in this pair is the *regression test.* Teaching that focusses on the provision of algorithms rather than improving students' conceptual understanding is referred to as *superficial teaching*. One decimal is said to be a *truncation* of another if it is created by truncating the longer decimal. For example, 4.45 is a truncation of 4.4502.

As discussed in Chapter 2, the term *repair* will be used in the sense of Brown and VanLehn (1982), "… many bugs can best be explained as patches derived from repairing a procedure that has encountered an impasse whilst solving a particular problem", (p122).

The next section of this chapter contains a discussion of various algorithms (both complete and incomplete) that enable a student to choose correctly on many decimal comparisons and a discussion of what can go wrong when students are applying these algorithms. Section 8.3 presents details of seven students who each regressed on two occasions. The issues to be investigated throughout this chapter will be highlighted using these students as examples. As in previous chapters, the analysis of the phenomenon is then conducted from two perspectives, the test-focussed approach (section 8.4) and the student-focussed approach (section 8.5).

8.2 Algorithms for comparing decimals

This section reviews the algorithms for comparing decimals that lead to students completing tests coded as A (A1, A2 or A3), because movement between the A codes is relatively common. Recall that a test is allocated the code A on correct answers to the 10 "core" items from Type 1 and Type 2 (see section 3.3.1). These are straightforward items without zeros or common digits. A test which has already been allocated the code A will then be allocated the code A1 if very few errors are made on the remaining test items from Types 3, 4, 5 and 6 (see Table 3.3). Such a student is referred to as a *task expert* to remind us that completing a test coded as A1 does not imply that the student has a full, rich and deep conceptual understanding of decimal notation. They may be using an algorithm with little understanding as discussed below. Previous misconceptions may still be held, masked by the expert performance of a learned algorithm.

A test coded as A will be allocated the code A2 if the errors are concentrated on items in Type 4 which involves decimals where one is a truncation of the other (for example 4.4502 / 4.45). In particular, the student makes at least 3 errors on these 4 items. Otherwise a test coded as A, but not meeting the criteria for either A1 or A2, is coded as A3.

Various algorithms for assisting students to complete tests coded as A are now discussed. Their similarities and differences will be illustrated with reference to some sample items and then suggestions are made for how students might react when their algorithm reaches an impasse.

8.2.1 Complete algorithms

As noted in Chapter 1, there are two algorithms that always provide the correct answer to any decimal comparison: *annexe zero algorithm* and *left-to-right digit comparison algorithm*. Both of these seem to be taught in the schools in our sample, although they are not both taught in all other schools systems (e.g. annexe zero is not taught in Japan, Dr Keiko Hino, University of Nara, personal communication). Evidence that students use the first algorithm is found on test papers; some students have written over 40 zeros on their test paper. Evidence that some students use the second was found in interviews with students; that these algorithms were sometimes being followed without any conceptual understanding was evident when students were unable to complete other decimal tasks (like insert a number between 0.3 and 0.4).

For finite decimal numbers with the same whole number portion, the algorithms are as follows:

Annexe zero algorithm

- Step 1: Write zeros at the end of the shorter decimal until both numbers have the same number of decimal places.
- Step 2: Treat the decimal portions as whole numbers.
- Step 3: The larger whole number (in the decimal portion) is the larger decimal.

Full left-to-right digit comparison algorithm

- Step 1: Moving from left to right, digits in corresponding columns are compared, until a difference is found.
- Step 2: If no difference in the digits is found, need to know that a *zero* can replace a *space*.
- Step 3: The first decimal with the larger digit is the larger decimal.

8.2.2 Incomplete algorithms

Use by students of the following incomplete algorithms was proposed by Steinle and Stacey (2001) to explain varying facilities of items within DCT1 and DCT2. These incomplete algorithms will work in many cases, but not all; see Table 8.1 for examples of items where they fail (as indicated by F). The incomplete algorithms below have been written to apply only to numbers with the same whole number portion.

Partial left-to-right digit comparison algorithm (PLR)

- Step 1: Moving from left to right, digits in corresponding columns are compared until a difference is found.
- Step 2: The decimal with the larger digit is the larger decimal, otherwise undecided.

Note how this algorithm differs from the full algorithm described above. Steinle and Stacey (2001) proposed that some students were using this *partial left-to-right digit comparison algorithm*, which fails to provide them with a definite choice in a very specific situation. For example, when comparing 8.24563 with 8.245 a student using this incomplete algorithm would compare 2 with 2, then 4 with 4, then 5 with 5 and then attempt to continue to the right by comparing 6 with a *space*. For some students, this is not a clear choice. The student needs to be able to *see the invisible zero* after the 5 in the shorter number to complete this strategy.

Round to 1 decimal place algorithm (R1)

- Step 1: If a decimal has more than one decimal place, *round* to one decimal place.
- Step 2: The decimal with the larger digit in the tenths column is the larger decimal, otherwise undecided.

Truncate to 1 decimal place algorithm (T1)

- Step 1: If a decimal has more than one decimal place, *truncate* to one decimal place.
- Step 2: same as for *Round to 1 decimal place*.

Round to 2 decimal places algorithm (R2)

- Step 1: If a decimal has more than two decimal places, *round* to two decimal places.
- Step 2: If a decimal has only one decimal place, annexe a zero.
- Step 3: Treat the decimal portions as the number of cents, number of centimetres or a percentage; in all cases, these are whole numbers between 0 and 99.
- Step 4: The decimal with the largest whole number (in the decimal portion) is the larger decimal, otherwise undecided.

Truncate to 2 decimal places algorithm (T2)

- Step 1: If a decimal has more than two decimal places, *truncate* to two decimal places.
- Steps 2 to 4: same as for *Round to 2 decimal places (R2).*

Note that both R2 and T2 might lead to *money thinking* (see section 3.2.3). The similarities and differences between these algorithms can be illustrated by considering some sample items; see Table 8.1. As well as the predictions for the various algorithms, this table contains the number of students who made errors on each item, given that the tests were allocated the coarse code A. This sample of 1902 tests was completed in 1997, when the DCT2 was introduced. While 1902 students were allocated the code A, exactly 1200 students made no errors on the test. So the remaining 702 students scored a maximum of 29/30 and contributed to the counts of errors in Table 8.1.

Table 8.1: Predictions for incomplete algorithms on DCT2 items

Item details (larger first)			Number of errors on A tests	Incomplete Algorithms PLR	R1	T1	R2	T2
	Type 1 items							
Q6	4.8	4.63	19	√	√	√	√	√
Q7	0.5	0.36	22	√	√	√	√	√
Q8	0.8	0.75	19	√	√	√	√	√
Q9	0.37	0.216	56	√	√	√	√	√
Q10	3.92	3.4813	45	√	√	√	√	√
	Type 2 items							
Q16	5.736	5.62	32	√	√	√	√	√
Q17	0.75	0.5	31	√	√	√	√	√
Q18	0.426	0.3	12	√	√	√	√	√
Q19	2.8325	2.516	14	√	√	√	√	√
Q20	7.942	7.63	11	√	√	√	√	√
	Type 3 items							
Q12	4.7	4.08	56	√	√	√	√	√
Q13	3.72	3.073	63	√	√	√	√	√
Q14	2.621	2.0687986	61	√	√	√	√	√
Q15	8.514	8.052573	51	√	√	√	√	√
	Type 4 items							
Q21	4.4502	4.45	201	F	F	F	F	F
Q22	17.353	17.35	157	F	F	F	F	F
Q23	8.24563	8.245	188	F	F	F	F	F
Q24	3.2618	3.26	178	F	F	F	F	F
	Type 5 items							
Q3	0.4	0.3	42	√	√	√	√	√
Q4	1.85	1.84	30	√	√	F*	√	√
Q5	3.76	3.71	37	√	√	F*	√	√
	Type 6 items							
Q26	0.42	0.35	15	√	√	√	√	√
Q27	2.954	2.186	11	√	√	√	√	√
Q28	0.872	0.813	15	√	√	F*	√	√
	Supplementary items							
Q1	0.457	0.4	223	F	√	F	√	√
Q2	1.3	0.86	91	√	√	√	√	√
Q11	1.06	1.053	62	√	F	F	√	√
Q25	3.746	3.741	31	√	F*	F*	√	F*
Q29	0.04	0.038	133	√	F	F	F	√
Q30	0.53	0.006	53	√	√	√	√	√
Number of items where algorithm fails				5	7	11	5	5
Possible classification				Any A	Any A	Any A	Any A	Any A

F: algorithm fails
F: equal length decimals, so expect students to choose correctly*
PLR: partial left-to-right digit comparison
R1(R2): round to one (two) decimal places
T1(T2): truncate to one (two) decimal places

While the two complete algorithms will provide the correct solution on each item, the five incomplete algorithms reach an impasse on different items (marked as F to indicate failure of the algorithm) and students then need to make a decision at this point; they need to *repair* their algorithm. Some students might then guess randomly, or might revert to a previously held misconception or to a currently held misconception that has been temporarily masked by use of the algorithm. In all cases, a different strategy (possibly guessing) needs to be implemented where an F is indicated.

Note that all of the five incomplete algorithms fail on the Type 4 items. Between 150 and 200 of the students in this sample made errors on the Type 4 items. A student using one of the incomplete algorithms needs to decide what to do when their algorithm fails. Steinle and Stacey (2003a) suggest that when their algorithm fails to produce a definite answer, that the repair may be to choose on the relative number of digits (i.e. L or S behaviours). If they consistently choose the *shorter* decimal, they will choose incorrectly on all of the Type 4 items and then the test will be coded as A2, although there are elements of S behaviour evident. A student who consistently chooses the *longer* decimal will choose correctly on all of the Type 4 items and then the test will be coded as A1. If they guess at random on the Type 4 items, they may choose correctly on 2 of the four items and hence the test would be coded as A3. (Note that this is not the only way for a test to be allocated the code A3.)

This section has detailed two complete and five incomplete algorithms for comparing decimals. Evidence for students using these algorithms will be sought throughout this chapter.

8.3 Case studies: students with two regressions

Before discussion of the 342 students who regressed from expertise at some time in this study, a special subset of seven students will be presented. These seven students regressed not just once, but twice! Table 8.2 lists their ID numbers and test histories; the *unstable A1 tests* are underlined and the *regression tests* are shaded.

Interestingly all of these students have their first unstable A1 test in Grade 7. This may just be a feature of the sample, as there needs to be at least 3 further tests. On the other hand, it may be that Grade 7 is special in some way. For example, as students move into their first year at secondary school, many teachers spend time revising material that they expect has been taught in primary school. If a student is unable to correctly order decimal numbers, it is possible that this revision assists them in the short term, but if a full understanding is not achieved, regression may be the long-term result. So, there might be a high rate of regression from an unstable A1 test in Grade 7, as the revision process is providing only superficial teaching to students with a lack of conceptual understanding. This will be investigated in section 8.4.1.

Table 8.2: Seven students with two regressions each

ID	Grade 6		Grade 7		Grade 8		Grade 9		Grade 10	
100701035				A1	U1			A1	U1	U1
300704079			A1	U1	A2	A1	U1			
390704012			L1	A1	U1	A1	S3			
400704005			A1	A2	A1	A2	A1			
410602014	A1		A1	U1	A1	U2	U1			
500703026				A1	A3		A1		A2	
600703029				A1		U1	A1	A1	A3	

Note: unstable A1 tests are underlined and regression tests are shaded

Regression to U1 occurred in 7 of the 14 regressions above, while there were no regressions to any of the L classifications and only one case to an S (i.e. S3, which is consistent with the hypothesis that interference of new teaching might cause students to use *negative thinking*; see section 3.2.2). Recall that a test is classified as U1 if it does not fit any of the expected patterns of choices that would be made by students with a misconception.

One student regressed to U2; this code has a low prevalence and hence its appearance in this table is unexpected. Section 3.2.4, following Steinle and Stacey (2002), suggested three possibilities for the code U2: *mis-read* the instructions (and therefore choose the smaller number in each item), *mis-chievous* (an expert who is having a bit of fun) and *mis-rule*. The code U2 will be analysed further in this chapter.

Student 400704005 is of interest as he oscillated between A1 and A2 on the 5 tests. As explained above, this could happen when a student is using one of the incomplete algorithms and when the algorithm reaches an impasse, the *repair* which is then applied to the algorithm is to choose the longer decimal (resulting in A1) on one occasion, and on another (later) occasion, the shorter decimal (resulting in A2). The five tests from this student were examined; Table 8.3 contains the responses by this student to the only five items that were answered incorrectly on any of the five tests. The last three rows indicate the total score out of 30, the score out of 4 (for the Type 4 items) and the code allocated to the test. A comparison of these responses to those in Table 8.1 indicates that these responses are entirely consistent with the use of one of the following incomplete algorithms: either *partial left-to-right digit comparison* (PLR) or *truncate to one decimal place* (T1). Brown and VanLehn predicted that movement between certain bugs may be an indication of which bugs are related as different repairs to the same incomplete procedure.

Table 8.3: Responses to various items on DCT2 by student 400704005

Item (larger first)			Type	Test 1 Gr7 Sem1	Test 2 Gr7 Sem2	Test 3 Gr8 Sem1	Test 4 Gr8 Sem2	Test 5 Gr9 Sem1
Q1	0.457	0.4	Supp	√	X	√	X	√
Q21	4.4502	4.45	Type 4	√	X	√	X	√
Q22	17.353	17.35	Type 4	√	X	√	X	√
Q23	8.24563	8.245	Type 4	√	√	√	X	√
Q24	3.2618	3.26	Type 4	√	X	√	X	√
Score out of 30 items				30	26	30	25	30
Score on Type 4 items				4	1	4	0	4
Code allocated to test				A1	A2	A1	A2	A1

Student 500703026 may also be consistently using one of the incomplete algorithms from Grade 7 to 10. Her first test was conducted in 1996 (TP3) and as she only made one error on the 25-item DCT1, this test was allocated the code A1. DCT2 was used for her other three tests and details of her responses (to the only five items that she answered incorrectly on any of the tests) are provided in Table 8.4.

Table 8.4: Responses to various items on DCT2 by student 500703026

Item (larger first)			Type	Test 2 Gr8 Sem1	Test 3 Gr9 Sem1	Test 4 Gr10 Sem1
Q1	0.457	0.4	Supp	X	√	X
Q21	4.4502	4.45	Type 4	X	√	X
Q22	17.353	17.35	Type 4	√	√	X
Q23	8.24563	8.245	Type 4	X	√	X
Q24	3.2618	3.26	Type 4	√	√	X
Score out of 30 items				27	30	25
Score on Type 4 items				2	4	0
Code allocated to test				A3	A1	A2

In summary, the examination of the test histories of students who regressed on two occasions has provided some supporting evidence for the claim that within the sample there are students with incomplete procedures who are choosing various repairs, which sometimes provide them with correct choices. Furthermore, it has generated these questions:

- Do students in Grade 7 have higher levels of regression than other students?
- Do the codes U1, U2, A2 and A3 appear more than expected as regression tests and if so, what does it reveal about these codes?

8.4 Test-focussed approach: consecutive test pairs

As mentioned above, the rate of regression from A1 was found to be 1 in 10 in Chapter 5. This was determined by using pairs of tests, where each student can contribute several times to the count. For example, student 410602014 (see Table 8.2) has a test history of [A1,A1,U1,A1,U2,U1] and would contribute on *three* occasions to the total number of test pairs where the first test in pair is A1. The first occasion is the pair (T_1, T_2), which is (A1, A1) and does *not* indicate regression. The second occasion is (T_2, T_3), which is (A1, U1) and *does* indicate regression. The final occasion is the pair (T_4, T_5), which is (A1, U2) and *does* indicate regression.

The total number of regressions using the test pairs approach was 349; see for example, row 1 of Table 5.17, which indicates that of the 3279 suitable test pairs, 2930 had the second test A1, which gives a difference of 349. Given the seven students in Table 8.2 who regressed twice, this indicates the actual number of students exhibiting regression at some stage in their test history is 342. For the rest of this section, the analysis and discussion relate to the 349 regressions while section 8.5 focusses on the 342 students involved.

8.4.1 Regression rates by grade

The case studies suggest that there may be more regression from an A1 test in Grade 7. In this section, the rate of regression for each grade will be investigated in order to determine if this is true. Note that there are two grades (or semesters) that are associated with a regression pair of tests. For the main analysis in this section, the grade of the first test will be used as this provides an indication of the unstable A1 tests at each grade. (For the analysis of the codes of the regression tests in the next section it is more appropriate to use the grade of the second test.)

Table 8.5 provides the grades of the two tests involved in the regression pair of tests; the numbers in the last row and last column will be used in analyses in this section. Grade, rather than semester, was chosen to provide numbers large enough to subdivide by school group and by the code of the regression test. (Table 1 in Appendix 10 provides details of the two semesters involved in each regression pair of tests, which shows, for example that 73% of the regression pair of tests were in consecutive semesters.)

Table 8.5: Numbers of regression test pairs by grades of involved tests

Grade of unstable A1 test	Grade of regression test						Total
	5	6	7	8	9	10	
4	3	1					4
5	10	16	1				27
6		23	18	1			42
7			28	74	7		109
8				36	67	3	106
9					10	43	53
10						8	8
Total	13	40	47	111	84	54	349

To investigate whether Grade 7 has a higher rate of regression, Table 8.6 provides the regression rates by the grade of the A1 test. The first column indicates the total number of test pairs that start with A1; the second column is the subset of these that retest as A1; the third column being the difference between the first two columns.

Table 8.6: Regression rate by grade of unstable A1

Grade of A1 test	Number of test pairs			Regression rate (%)
	Total (A1,X)	Retest (A1,A1)	Regression (A1,non-A1)	
Grade 4	10	6	4	40
Grade 5	178	151	27	15
Grade 6	512	470	42	8
Grade 7	1057	948	109	10
Grade 8	953	847	106	11
Grade 9	449	396	53	12
Grade 10	120	112	8	7
Total	3279	2930	349	11

Note that the regression rate for students in Grade 7 is not higher than the other grades, as was predicted from the case study of seven students earlier this chapter. In fact, with the exception of Grade 5, the regression rate is fairly constant across grades. As the sample size is only 10 in Grade 4, a more reliable figure is obtained by grouping the two youngest grades; students from Grades 4 and 5 have an overall rate of 16% regression. Recall that a negative project effect was found for students in Grade 5; see section 5.2.2. If Grade 5 students who were completing their first test were provided with superficial teaching, then the prevalence of expertise would be higher than expected and this would generate a negative project effect when compared with other Grade 5 students. Such superficial teaching would also explain the high rates of regression from A1 tests in Grade 5.

An alternative approach, using the grade of the regression test, (details provided in Table 2 in Appendix 10) confirms that there is a higher rate of regression for the youngest students, (28% in Grade 5, n=46), and that the remaining grades have a flat distribution of 11%, (with the exception of a low of 7% in Grade 7). Hence, if there is any superficial teaching before the testing was conducted, then it appears to have been confined to Grades 4 and 5; Grade 7, in particular, does not have a high rate of regression.

8.4.2 Regression rates by school group

Table 8.7 provides similar details to Table 8.6 but this time the test pairs are grouped by school group rather than grade. It is clear that all school groups are involved in regression to some extent; SGE has the lowest regression rate of 5% and SGA the highest (16%). It had been proposed that the school group with the highest levels of expertise (SGE) might have accompanying high rates of regression, which would indicate that superficial teaching took place before the testing was conducted. As can be seen from this table, SGE had the lowest rates of regression, so there is no evidence of superficial teaching to artificially elevate the levels of expertise. It is not possible to discount, however, the possibility that superficial teaching was provided before each test was conducted. Similarly the regression rates for SGD, the school group where teachers looked at the DCT results as a league table, appear to be no higher than others.

Table 8.7: Regression rate by school group

School Group	Number of test pairs			Regression rate (%)
	Total (A1,X)	Retest (A1,A1)	Regression (A1,non-A1)	
SGA	327	275	52	16
SGB	413	376	37	9
SGC	698	630	68	10
SGD	618	546	72	12
SGE	615	582	33	5
SGF	608	521	87	14
Total	3279	2930	349	11

Before moving to the next analysis, note that Table 3 in Appendix 10 contains a cross-tabulation of the 349 regressions by the grade of the A1 test and school group.

8.4.3 Regression rates by test number

While various analyses were conducted (for example, by cohort) to investigate various factors that might explain variations in rates of regression, only one is presented below. Table 8.8 contains the rate of regression by considering the test numbers of the two tests involved. For example, there were 139 regressions that occurred from an A1 on a student's first test, leading to a regression rate of 15%. This is much higher than the regression rate from other test numbers (210 out of 2357 is 9%). One possible reason for this is that some (or all) school groups were involved in superficial teaching before the first testing was conducted. This will be investigated in the next section.

Table 8.8: Regression rate by test number

Test numbers	Number of test pairs			Regression rate (%)
	Total (A1,X)	Retest (A1,A1)	Regression (A1,non-A1)	
(T_1,T_2)	922	783	139	15
(T_2,T_3)	957	854	103	11
(T_3,T_4)	780	716	64	8
(T_4,T_5)	449	416	33	7
(T_5,T_6)	138	128	10	7
(T_6,T_7)	33	33	0	0
Total	3279	2930	349	11

8.4.4 Localised high regression rates from first tests

While the overall rate of regression is 11%, the three school groups with the highest rates were investigated with a focus on student's first and second tests (due to the findings in Table 8.8). Three particular rates of 30% or more were found:

- In SGA, there were 44 students who tested as A1 in Semester 2 of Grade 7 in TP5 (and had another test); 13 students (30%) regressed.
- In SGD, there were 23 students who tested as A1 in Semester 1 of Grade 5 in TP4 (and had another test); 7 students (30%) regressed.
- In SGF, there were 35 students who tested as A1 in Semester 2 of Grade 8 in TP3 (and had another test); 14 students (40%) regressed.

Hence, when a very detailed analysis is undertaken, higher levels of regression can be found, which may be evidence that particular classes have received superficial teaching before the test was conducted (in particular, the testing when the school joined the study). The data analysis used in this study does make this teaching effect difficult to pinpoint: superficial teaching will happen at the *class* level, possibly at the *school* level; but the general data analysis groups schools together. Effects due to specific teaching practices are therefore smoothed out.

There are also many examples of very low regression, but these are not reported. Overall, there are no grades that have consistently higher rates of regression with the exception of the students in Grades 4 and 5. High rates of regression are localised.

8.4.5 Codes of the regression tests

The discussion in this section has so far been on the rate of regression but gave no information on the code of the regression test, that is, which non-A1 test follows the unstable A1. The case studies raised the possibility that the codes A2, A3, U1 and U2 may be more often involved than the other codes. Is this true, and is this affected by the grade of the student?

Table 8.9 contains the distribution of the codes of the regression tests given the grade of the regression test; (Table 4 in Appendix 10 provides the number of regressions). Note that for this discussion, the grade of the second test in the pair is used, rather than the grade of the unstable A1.

Table 8.9: Distribution (%) of the regression test codes by grade of regression test

Grade of regression test	Codes of the regression tests									
	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Gr 5 (n=13)	23	0	0	15	0	0	8	0	54	0
Gr 6 (n=40)	18	10	5	15	0	10	8	0	33	3
Gr 7 (n=47)	21	13	2	6	0	11	0	4	43	0
Gr 8 (n=111)	31	12	3	3	0	4	5	11	24	8
Gr 9 (n=84)	23	17	0	5	1	5	4	5	32	10
Gr 10 (n=54)	26	19	2	2	0	2	6	2	33	9
Overall (n=349)	25	13	2	5	0	5	5	5	32	7

The most likely code involved in regression is U1 (32% overall) and the next most common code is A2 (25%). Regression to A3 is uncommon in younger students but has an increasing trend with grade. Students who regressed to A2 or A3 have been discussed earlier as likely to be using one of the *incomplete algorithms* that fail in certain comparisons. Movement between A codes can be explained by bug migration (different repairs to incomplete algorithms). While it appears that the rates to A2, A3, U1 and U2, are somewhat high, and the rates to the other codes are low, it may be that the students are regressing to all codes in the proportion that they appear in the general population.

To determine whether the codes for the regression tests are distributed in a similar manner to the general prevalence of the fine codes at each grade level, the *rescaled* distribution of non-A1 test codes at each grade level (see Table 5.5) was used in the following graphs. The codes are grouped according to whether the distributions of the codes involved in regression are typically above or below the general prevalence. Figure 8.1 contains the codes that are *over-represented* in regression (i.e. A2, A3, U1 and U2), while Figure 8.2 contains the codes that are *under-represented* in regression (i.e. L1, L2, L4, S1, S3 and S5).

These graphs confirm that A2, A3, U1 and U2 are over-represented in regression, while the L and S codes are under-represented. Within the six graphs in Figure 8.2, there are a few grades where the general relationship does not hold. For example, in Grades 5 and 6, there are more regressions to L2 than expected, although there are only 13 students with regression in Grade 5. Although the difference is slight, there are more regressions to S1 in Grades 6 and 7 than might be expected.

There is a definite peak, however, of 11% for S5 in Grade 8, (can be seen in Table 8.9) with a sample of over 100 students. This peak is accentuated by the fact that the other grades have fewer regressions than the general prevalence (see Figure 8.2f). This peak is consistent with the hypothesis that new teaching can interfere with a student's current understanding.

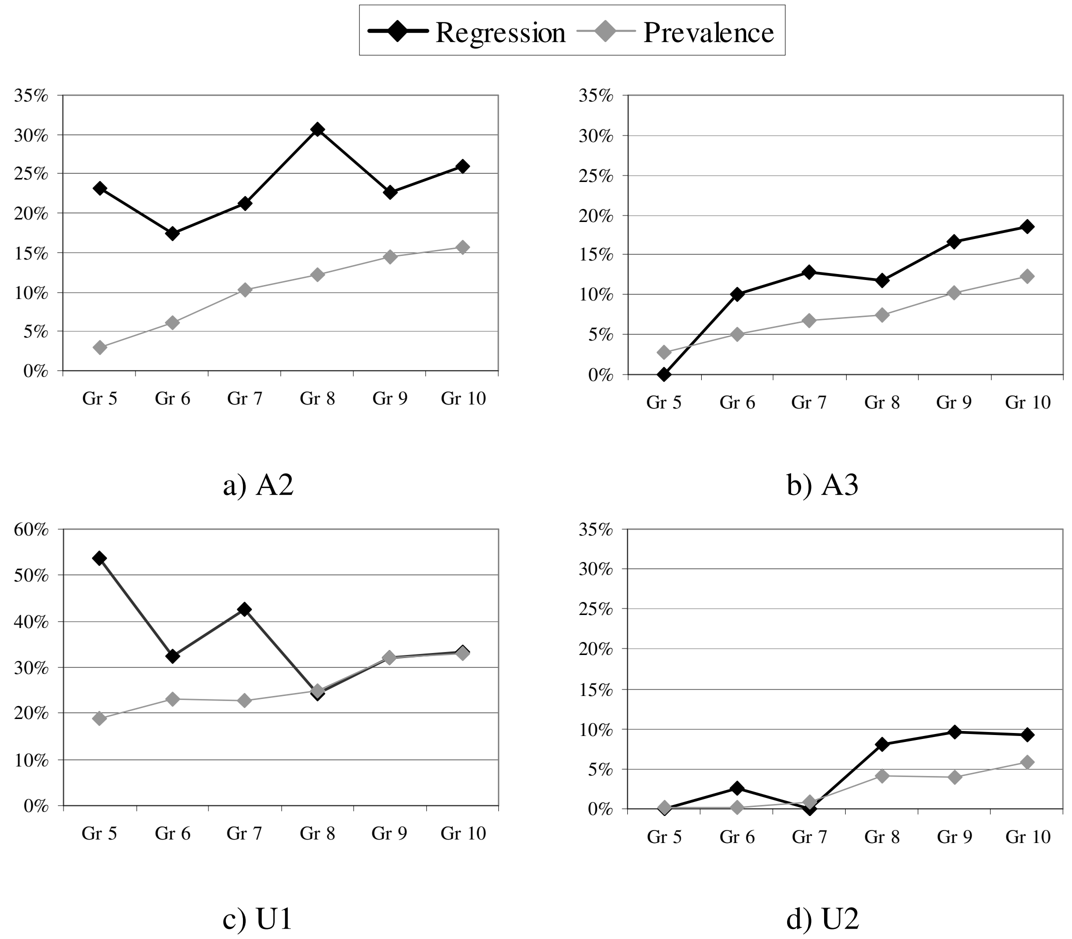


Figure 8.1: Comparison of distribution of codes involved in regression with general non-A1 prevalence, for codes that are over-represented in regression. (Note the different scale for the U1 graph)

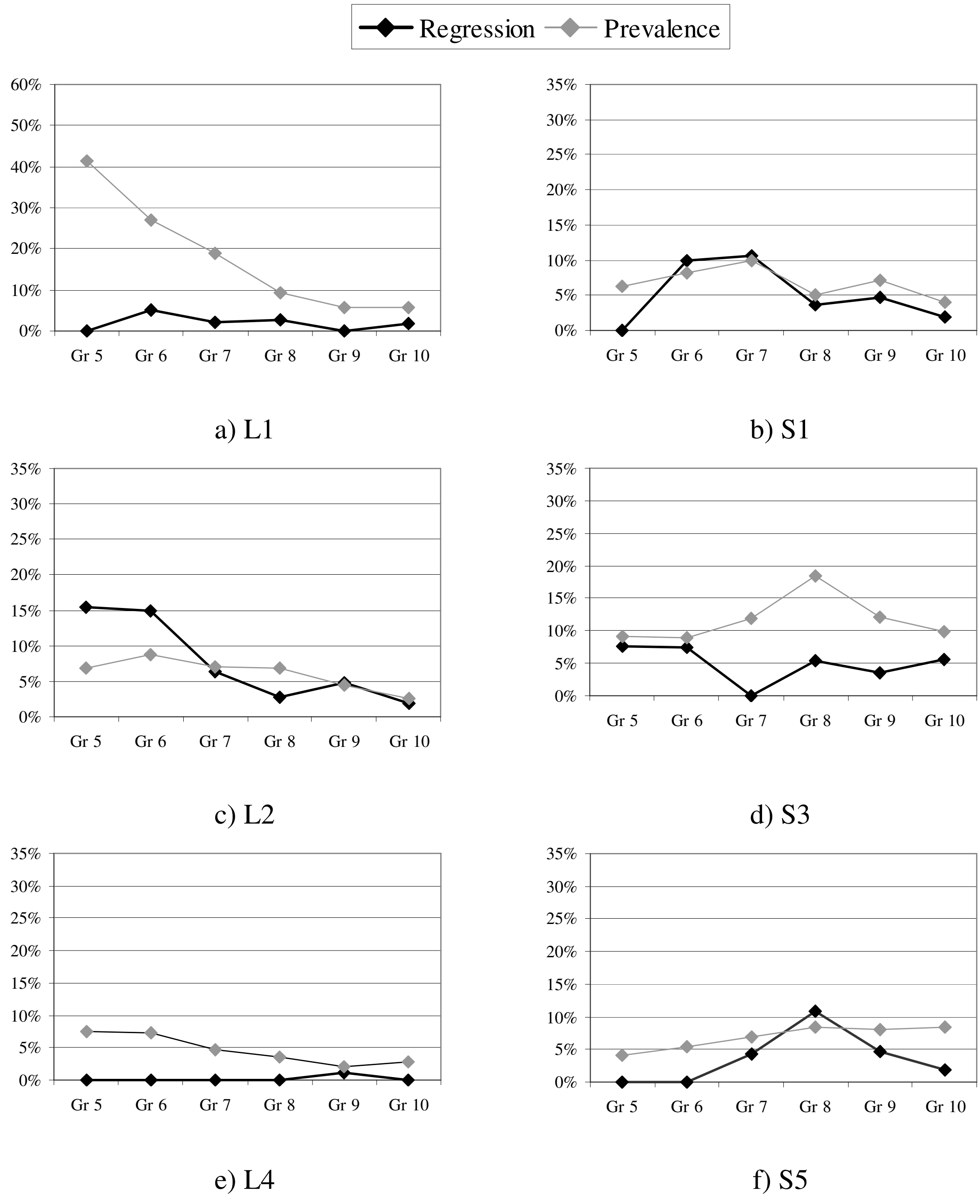


Figure 8.2: Comparison of distribution of codes involved in regression with general non-A1 prevalence, for codes that are under-represented in regression. (Note the different scale for the L1 graph)

To determine if there are variations to these results in the different school groups, the codes involved in regression were again subdivided, this time by school group instead of grade. Table 8.10 provides the distributions of these grouped codes by school group (see Tables 5 and 6 in Appendix 10 for numbers of students).

Table 8.10: Distribution (%) of grouped regression test codes by school group

School Group	Grouped regression test codes			
	A2/A3	Any L	Any S	Any U
SGA (n=52)	38	8	15	38
SGB (n=37)	43	3	27	27
SGC (n=68)	32	9	19	40
SGD (n=72)	35	7	15	43
SGE (n=33)	33	18	6	42
SGF (n=87)	46	6	10	38
Overall (n=349)	38	8	15	39

This table confirms that the L codes are unlikely to be involved in regression; the only school group with more than 10% to L was SGE with a sample of only 33 students involved in regression. About 15% of the regressions are to S, with a higher rate for SGB. This school group was noted earlier to have a higher prevalence of S in general; nearly 30% of the tests in the primary school were allocated the code S, which is double the overall figure, (see Table 5.3). It is also interesting that the school group with the highest regression to A2 and A3 (SGF) was noted earlier as having the highest prevalence of A2 and A3 (see Table 5.4).

(It is appropriate to note that the codes of the regression tests are already counted in the test-focussed prevalence of each code calculated in Chapter 5. Considering the prevalence of the codes on students' first tests, however, confirms this same result for SGB and SGF).

Hence, while the distribution of the codes involved in regression within the school groups confirms the general trends, it also highlights that there are some particular teaching effects in two of the school groups.

The above analyses provided an overall rate of regression from one test to another of about 10%, and indicated that while there was some variation between school groups, students from all school groups were involved in regression at fairly similar rates. About 40% of regressions were to the codes A2 and A3, which is more than would be expected from the prevalence of the non-expert codes. While only 7% of regressions were to U2, this is about twice what would be expected from the prevalence of the non-expert codes. This provides additional information about the codes A2, A3 and U2, which will be consolidated in Chapter 9.

8.5 Student-focussed approach

As in earlier chapters, another analysis can be undertaken which focusses on students rather than the pairs of tests that they generate. This alternative analysis will now be discussed.

8.5.1 Regression rates

Another approach to calculating the rate of regression is to consider each student; of the 3204 students in the study, 2137 tested as A1 at some stage in their test history. Some of these students (n=595), however, had their first A1 on their last test and so need to be excluded from the sample, as they had no possibility of regressing. This leaves 1542 students who had the potential to regress as they had between 1 and 6 extra tests after their first A1. (In fact, the average number of extra tests for these 1542 students is 2.3.) As mentioned above, the actual number of students who regressed was 342, which is 22% of the 1542 students who had the opportunity. These figures are summarised in Table 8.11, (see last column) as well as details by school group.

Table 8.11: Regression rate by school group using student-focussed approach

Description	School Group						Overall
	SGA	SGB	SGC	SGD	SGE	SGF	
Total number	709	679	549	397	258	612	3204
Number with some A1	350	452	357	291	228	459	2137
Student prevalence of A1	49%	67%	65%	73%	88%	75%	67%
Potential students*	217	265	290	247	200	321	1542
Number who regress	51	37	66	70	32	86	342
Regression rate	24%	14%	23%	28%	16%	27%	22%

**exclude students whose first A1 is on their last test as they cannot possibly regress*

As noticed in the previous section (using the test pair approach), SGE has a relatively low rate of regression compared to the other school groups and the highest prevalence of expertise. SGA has the opposite characteristics: higher rates of regression and the lowest prevalence of expertise.

This new rate of regression is double the rate found using the test pair approach. This seems reasonable when compared with theoretical calculations of probability of regression assuming at least two tests person (average is 2.3) and probability of regression on one test of 0.11.

For example:

Pr(regression within two tests after A1)
=Pr(regression on the first test after A1)
+Pr(no regression on first test after A1, then regression on the second test)
=0.11 + 0.89 x 0.11
=0.21

This model assumes that regression is as likely after a second A1 test as after the first. It predicts that had students been tested more times, regression would increase. For example, with three tests, it predicts a regression rate of just less than 0.3 (0.11+0.89 x 0.11+0.89 x 0.89 x 0.11= 0.295).

So we can consider the regression rate either as 1 in 10 (by considering regression at the next test), or nearly 1 in 5 (eligible) students in the study; this second figure being much more dramatic. This confirms that regression is not an isolated phenomenon, but occurred in all school groups and affected one in five students at some time in this study. Noting the earlier increase in regression rates from test pairs by SGA, SGD and SGF on student's first tests, (section 8.4.4) it might be reasonable to expect that the corresponding regression rates by students in these school groups are likewise higher. Removing these three school groups leaves another three school groups with rates of 14%, 16% and 23%. Note that, of these, SGB had the least regression (14%) and was amongst the less-tested school groups (average of less than three tests per student) and SGC has the most regression (23%) and was amongst the more-tested school groups (average of more than four tests per student). Hence, as was noted in Chapter 5, students with less testing dilute the student-focussed measures as they have fewer opportunities to exhibit certain characteristics.

8.5.2 The test preceding the unstable A1 test

The test before the unstable A1 test may provide information about whether students are regressing to their earlier way of thinking. In this case, the temporary expertise most likely occurred as a result of superficial teaching. Interference from new teaching is another possible reason for regression, although this may move students to a different code and will not be easily detected.

While the numbers of students with particular combinations of codes are contained in Table 7 in Appendix 10 (for the test *before* the regression pair), the distributions obtained by grouping the fine codes into coarse codes are provided in Table 8.12. Of the 342 students with regression, 139 had their regression pair as their first 2 tests and therefore no previous test. This leaves 203 students with at least one test before the unstable A1 test. The last column of Table 8.12 indicates that 36% of these 203 students (73 students) had an A1 before the unstable A1, so their test histories contain (A1, A1, non-A1).

The shaded diagonal cells in Table 8.12 highlight the tendency for students to regress to the same code as the test preceding the unstable A1. For example, of the 28 students who regressed to S (and who had at least one test preceding the unstable A1), 36% were S on this earlier test.

Table 8.12: Distribution (%) of earlier test code, given regression test code

Test preceding the unstable A1	Code of regression test				Overall (n=203)
	A2/A3 (n=87)	L (n=16)	S (n=28)	U (n=72)	
A1	43	19	18	39	36
A2/3	18	6	4	8	12
L	11	56	18	13	16
S	13	6	36	15	16
U	15	13	25	25	20
Total	*100*	*100*	*100*	*100*	*100*

Overall, of the 130 students who did not have an A1 immediately preceding the unstable A1, 32 students (25%) regressed to the same fine code. If the coarse codes are used, (this requires A2 and A3 to be grouped) then there were 53 students (41%) who returned to the same coarse code as before the unstable A1 test. Hence, there is some evidence of students regressing to their previous code that is consistent with superficial teaching.

8.6 Conclusion

The overall rate of regression from A1 is about 10%, determined by considering consecutive pairs of tests. Younger students have slightly higher rates. However, a very different picture is obtained if individual students are considered. Of the students with the opportunity to regress (i.e. they had another test after their first test as experts) about 20% regressed at some time in their test history. So, 1 in 5 students who appeared to be experts, tested as non-expert in a later test, indicating that these students did not have a robust understanding of decimal notation.

Several situations that might lead to regression were proposed and there is some evidence in the data to support these suggestions. For example, several school groups were shown to have higher rates of regression associated with joining the study, (section 8.4.4), which is consistent with superficial teaching before the first testing took place.

Some codes, in particular A2 and A3, were over-represented in regression. This is consistent with the hypothesis that there are students using *incomplete* algorithms that allow them to make correct decisions on the more commonly met decimal comparisons. When presented with certain comparisons (such as 17.353 / 17.35), their algorithm fails to provide an answer and as they have no conceptual understanding to then assist them, they are unsure of what to do. In particular, students with a mixture of the codes A1, A2 and A3 confirm that students are using these incomplete algorithms. We believe that particular teaching techniques, such as always rounding decimals to two decimal places, or an over-reliance on money as a context for decimals may reinforce these wrong ideas in students' minds and not assist students to see the generalised structure of our number system. Whether the students are referring to money, percentages or another context to two decimal places, and whether they round or truncate is often irrelevant. The main point is that the lack of understanding of decimal notation (in particular of expanded notation i.e. that 17.352 is 17 ones + 3 tenths + 5 hundredths + 2 thousandths and hence *must be* larger than 17 ones + 3 tenths + 5 hundredths) is resulting in students memorising apparently meaningless algorithms that may only be partially remembered and/or misapplied.

The rate of regression to S5 in Grade 8 was about double the rate in any other grade. This is consistent with interference of students' understanding by new teaching, as mentioned in Chapter 2. In particular, it was suggested in section 3.2.2 that attempts by students to assimilate negative numbers into their existing number system may lead to S behaviour as might the introduction of students to very small numbers using scientific notation, ($0.000064 = 6.4 \times 10^{-5}$).

It was proposed that the school group with the highest levels of expertise (SGE) may have accompanying high rates of regression, but this was shown not to be the case. There is no evidence that teachers in this school group were artificially inflating their students' results; either they taught this skill of decimal comparison well, or they were always diligent at revising the task before the testing took place.

CHAPTER 9 FROM CODES TO THINKING

9.1 Introduction and terminology

Each of the analyses already reported in Chapters 5 to 8, as well as several new analyses that will be introduced in this chapter, provide information about the codes and the related ways of thinking. These will be synthesised in this chapter to provide a full picture of the codes.

While both the test-focussed prevalence (TFP) and student-focussed prevalence (SFP) of the coarse and fine codes were determined in Chapter 5 by a straightforward interrogation of the data, the improved estimates presented in this chapter overcome some of the limitations of the dataset. For example, due to the project effect (higher levels of expertise by students who have completed earlier tests compared to students in the same grade who have not), only students' first tests will be used to determine the test-focussed prevalence of each code at each grade. While these results are provided graphically within this chapter, Table 19 in Appendix 4 contains the relevant percentages.

On the other hand, calculations of the student-focussed prevalence of the codes were being diluted by students who were tested on few occasions as such students have less opportunity to contribute to several different codes. Hence, the less-tested students were removed to create two samples of students with at least four tests in either primary school or secondary school. These two samples were then used to determine the improved student-focussed prevalence of the codes within the two school levels. Results only are provided in this chapter (full details in Appendix 5).

Terminology

Students' test histories have been used in various analyses throughout this thesis; ignoring all occurrences of U tests now creates *condensed test histories*. For example, a student with a test history of [L,U,S,U,A,A] has a condensed test history of [L,S,A,A].

The *association* of code X with code Y is the proportion of students with code X (and a total of two or more tests) who also have code Y in their test history. Note that the association of X with Y is unlikely to be the same as the association of Y with X.

Persistence-proximity maps are two-dimensional graphs created by combining two measures from Chapters 6 and 7. The X-axis represents the persistence of students in a code and the Y-axis represents its proximity to expertise (the rate of movement to A1 on the next test, given that the student changes code, i.e. does not persist). Tables 2 to 11 in Appendix 9 have the sample sizes for the data points in these figures.

Section 9.2 presents these three analyses (condensed test histories, association between codes and persistence-proximity maps), which are designed to summarise and extend the findings from earlier chapters.

The fine codes will then be discussed in turn, and are grouped (for convenience) into coarse codes A, L, S and U (Sections 9.3 to 9.6). Within each section, the improved prevalences of the codes (both test-focussed and student-focussed) are presented. These are followed by graphs indicating the distribution of the second and third tests based on the condensed test histories (similar to Figure 9.1) and then persistence-proximity maps (similar to Figure 9.2, but split by grade). The discussion of the fine codes that follows then draws on this information as well as the measure of association between the fine codes in Table 9.2. Finally, the new insights gained from the analysis of the longitudinal dataset into the ways of thinking that lie behind the codes are discussed in detail.

9.2 New analyses

9.2.1 Condensed test histories

The aim of this section is to consider pathways that students follow through the misconceptions. For each student, ignoring all occurrences of U tests created a condensed test history. This substantially reduces the number of possible sequences of codes while keeping all the important information. For example, if a sequence of three tests were to be examined, then using the 3 coarse codes (A, L and S) results in only 27 possibilities, compared with 64 if all 4 coarse codes are used. This is gauged to be a reasonable procedure as U is not so much a misconception, as an absence of known misconceptions. Furthermore, it was shown in section 5.4.1 that the code U has the highest turnover of the coarse codes.

A sample of 1663 students, who meet the criteria of at least three tests in their condensed test history, was created. Table 9.1 contains the distribution of codes (both coarse and fine) on the first test within these condensed test histories. There were 857 students with A on their first test (in their condensed test history), 521 students with L and 285 students with S. A decision was made to limit the following analysis to the first three tests in these condensed test histories, as later tests are more likely to be A1 due to the project effect. Appendix 6 contains various details of these 1663 students, such as the grades of the first and third tests in these condensed test histories.

Table 9.1: Distribution of first tests in students' condensed test histories (n=1663)

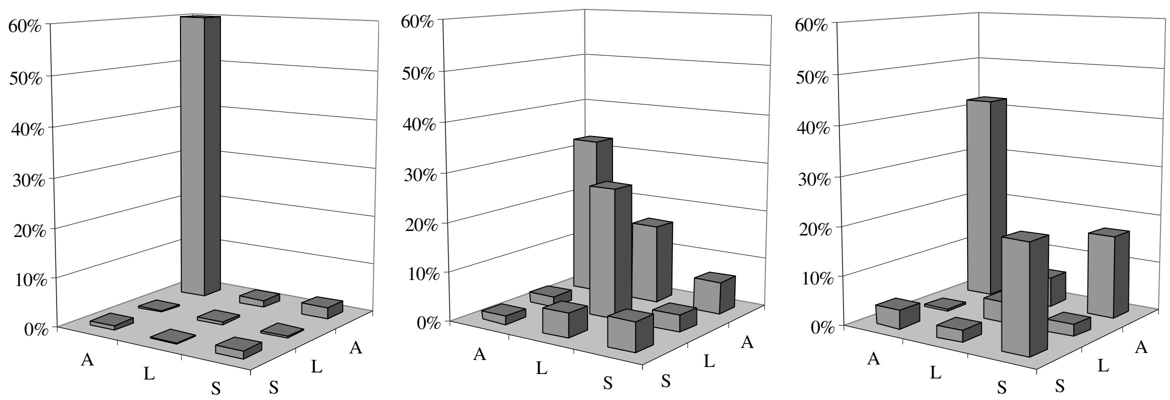
Coarse code	Tests		Fine code	Tests	
	Number	Percentage		Number	Percentage
A	857	52%	A1	733	44%
			A2	86	5%
			A3	38	2%
L	521	31%	L1	375	23%
			L2	70	4%
			L4	76	5%
S	285	17%	S1	88	5%
			S3	121	7%
			S5	76	5%
Total	1663	100%	Total	1663	100%

The graphs in Figure 9.1 are created from the distributions of the second and third tests in these condensed test histories, given the first test is either A, L or S; hence the sum of the columns in any one graph is 100%. The left side corresponds to the second test in the condensed test histories, and the right side to the third such test.

The tall column at the back of Figure 9.1a (truncated from 92%) indicates that it is unlikely that students whose first non-U test was A later complete tests coded as L or S. For students whose first non-U test was L, the column in the AA position (see Figure 9.1b) is much lower (less than 35%), and the central tall column (about 25%) corresponds to retesting as L on both the second and third non-U tests. For students whose first non-U test was S, the column in the AA position in Figure 9.1c is a little taller than for the L students just discussed, (over 40%), and the next tallest column (just over 20%) corresponds to retesting as S on both the second and third non-U tests.

From Figure 9.1 we can observe the higher rates to A from the S students in comparison to the L students, which is consistent with the finding that S students are above L students in a hierarchy (as discussed in Chapter 7). We also observe the general persistence by both L and S students (as discussed in Chapter 6).

As the focus of this chapter is the fine codes, similar graphs are provided for each fine code in the following sections. The most useful analysis is a hybrid of fine and coarse codes; the later graphs thus provide the distribution of the coarse codes for the second and third tests (in the condensed test histories) given the fine code of the first test.



a) A (n=857) b) L (n=521) c) S (n=285)

Figure 9.1: Distribution of second (left-hand axis) and third (right-hand axis) tests in condensed test histories, given that the first test is A, L or S. (Note that the AA column in (a) is truncated from 92%)

9.2.2 Association between the codes

Another new analysis that will be used in this chapter is referred to as *association*, and provides an indication of which pairs of codes may be linked as they appear more frequently within students' test histories. This was determined by considering a sample of students who were involved in a particular code, and then determining what proportion of these students were also involved in each other code.

The results for the fine codes are summarised below in Table 9.2, which gives the percentage of students (with at least two tests) involved in each other code. For example, while there were 2137 students with an A1 code (see Table 5.6), exactly 340 students had only one test and so could not possibly be involved in other codes. Of the remaining 1797 students involved in A1, Table 5.14 indicates that 213 students were involved in A2 tests at some stage and 149 were involved in A3 tests. These results are listed in the first row of Table 9.2; 12% of the A1 students (with at least two tests) were involved in A2 tests at some time in their test history, (213 out of 1797), while 8% of this same sample of A1 students was involved in A3. (There are no relations between the entries in the table: potentially all entries could be 0% if students never changed code and all would be 100% if all students had tests in every code.)

Inspection of Table 9.2 indicates that the largest numbers in a row tend to be under A1 and under U1; so many codes have higher association with A1 and U1 than other codes. Evidence from Table 9.2 will be used throughout this chapter; for example, compared with A1 and A2, A3 is more highly associated with L1.

Table 9.2: Association between pairs of fine codes*

Fine code		Fine code										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
A1	(n=1797)		12	8	16	7	5	7	8	7	23	3
A2	(n=307)	69		13	12	7	7	13	13	11	22	3
A3	(n=239)	62	17		28	10	9	10	12	8	37	5
L1	(n=580)	49	7	12		17	15	11	12	8	38	1
L2	(n=235)	53	10	10	41		12	11	12	7	37	1
L4	(n=191)	51	10	12	46	15		8	13	9	34	2
S1	(n=242)	55	16	10	26	11	7		23	20	33	2
S3	(n=336)	44	12	9	21	9	7	16		18	36	4
S5	(n=241)	49	14	8	19	7	7	20	25		39	5
U1	(n=728)	56	9	12	30	12	9	11	16	13		3
U2	(n=71)	66	14	15	6	4	4	6	17	15	32	

**proportion of students with row code who are also involved with column code (students with only one test are excluded from calculation)*

9.2.3 Persistence-proximity maps

Other diagrams that will be used in this chapter are created from the results in Chapters 6 and 7 (details provided in Appendix 9). Combining the measures of *persistence* (rate of retesting in the same code on the next test) and *proximity to A1* (rate of moving to A1, given a change of code, i.e. no persistence) provides a map of the relative positions of the codes in a two-dimensional space. (Note that using a conditional probability for the proximity to A1 ensures that the measures on the X- and Y-axes are independent; otherwise, as one measure increases, the other measure would be forced to decrease.)

Figure 9.2a provides this two-dimensional space of persistence against proximity to A1 for the tests conducted in the primary sample, while the corresponding result for the secondary sample is in Figure 9.2b. The codes are plotted with the X-axis representing the persistence (from Chapter 6) so the further the code to the right, the higher the persistence. The Y-axis represents the proximity to A1 (from Chapter 7), so the higher the code, the more likely it is that the student will move to A1 on their next test. An appropriate image might be that, in moving from the state of a novice to that of an expert (i.e. moving up the vertical axis), students are sometimes side-tracked; the length of this side-track (the horizontal distance out to the code) is an indication of the time that some students take to return to the main path. Codes that are plotted further to the right indicate the undesirable situation of students who are "stuck" in a code.

These graphs provide a map of the relative location of students in various codes in each school level. For primary students, the S codes are closer to A1 than the L codes, and the code that is the hardest for students to leave (furthest to the right) is L1 (39%). For secondary students, however, both S1 and S3 have dropped so that they are now *further* from expertise than they were in primary. In fact, these two codes are now even further from A1 than L1 is. Furthermore, persistence in S3 has increased, so that S3 has the highest persistence of all the codes in secondary school.

The fine codes will now be discussed in turn, and are grouped (for convenience) into coarse codes A, L, S and U (Sections 9.3 to 9.6). Within each section, the improved prevalences of the codes (both test-focussed and student-focussed) are presented. These are followed by graphs indicating the distribution of the second and third tests based on the condensed test histories (similar to Figure 9.1) and then persistence-proximity maps (similar to Figure 9.2, but split by grade). The discussion of the fine codes that follows then draws on this information as well as the measure of association between the fine codes in Table 9.2.

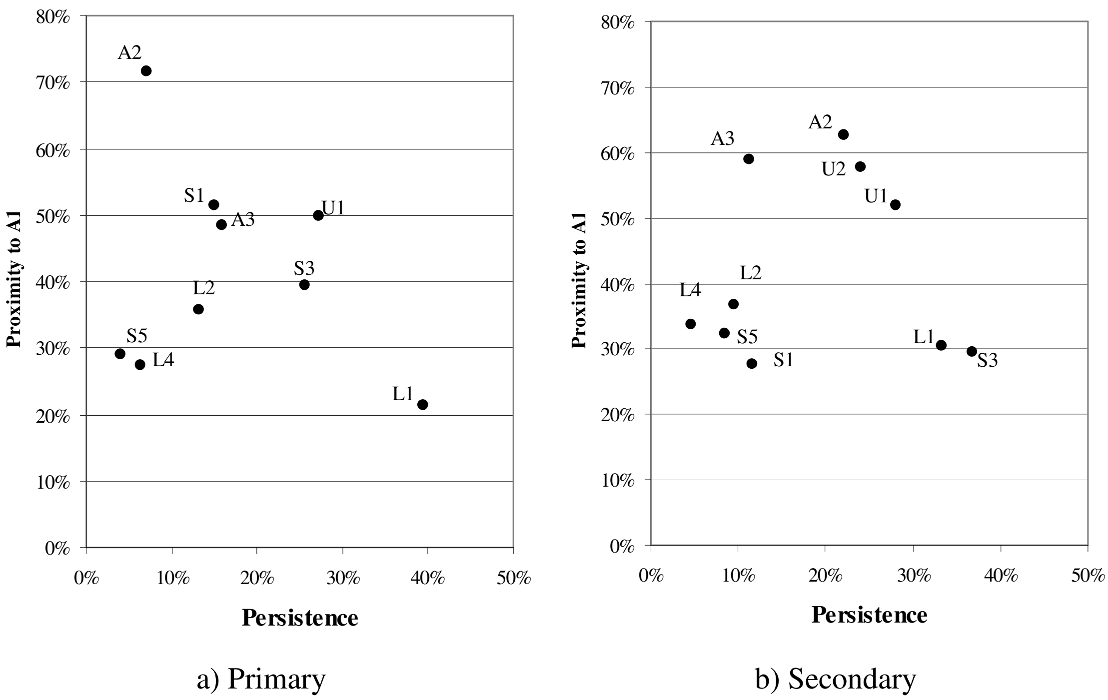


Figure 9.2: Persistence-proximity map for primary and secondary samples

9.3 The A cluster of codes

Figure 9.3 provides the improved test-focussed prevalence of A1, A2 and A3 by grade based on students' first tests. It clearly shows the dominance of the code A1 within A, and the increasing trends with grade. Note that the initial calculations for Figure 9.3 suggested a drop in the prevalence of A1 from Grade 7 to Grade 8, as the high performing SGE did not contain any students who completed their first tests in Grades 8, 9 or 10. To overcome this sampling issue, the prevalence of A1 has been adjusted to approximate figures that reflect the increasing prevalence in Figure 5.1.

Table 9.3 provides the improved student-focussed prevalence of these codes for each school level. Rather than provide the figures for A2 and A3 separately, the percentage of students who were involved in either code is provided. This is due to the finding in Chapter 8; the over-representation of both A2 and A3 in the tests following A1 tests indicates that these students are using incomplete algorithms. Note that 1 in 4 secondary students are expected to complete tests coded as either A2 or A3, that is, tests which are allocated the coarse code A but not A1.

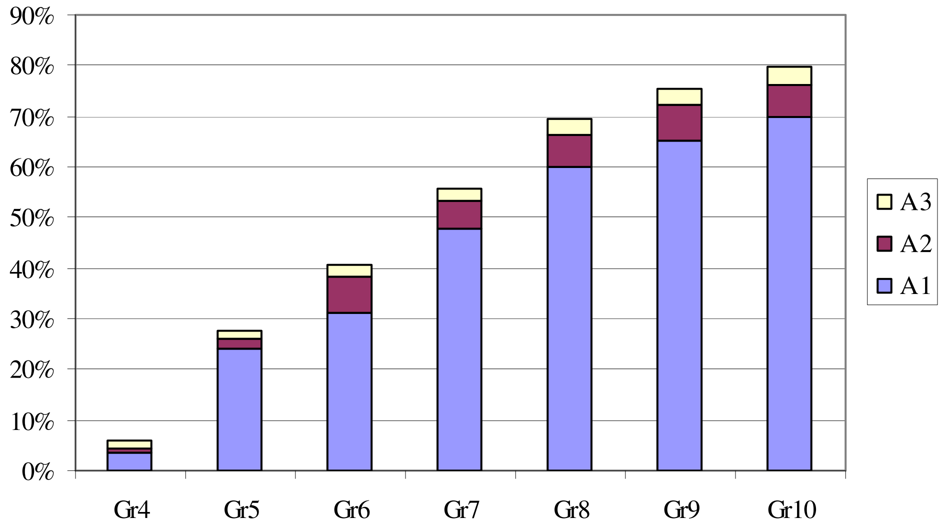


Figure 9.3: Improved TFP of A1, A2 and A3 by grade

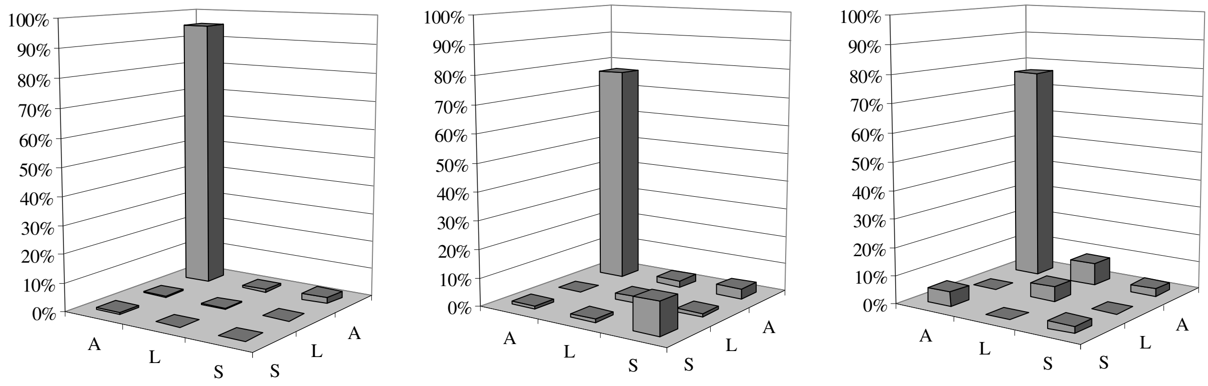
Table 9.3: Improved SFP of A, A1 and A2/A3 by school level*

School Level		A	A1	A2/3
Primary	(n=333)	68 (40)	60 (30)	17
Secondary	(n=682)	88 (80)	83 (70)	26

** Best estimates in brackets after adjusting figures to remove project effect*

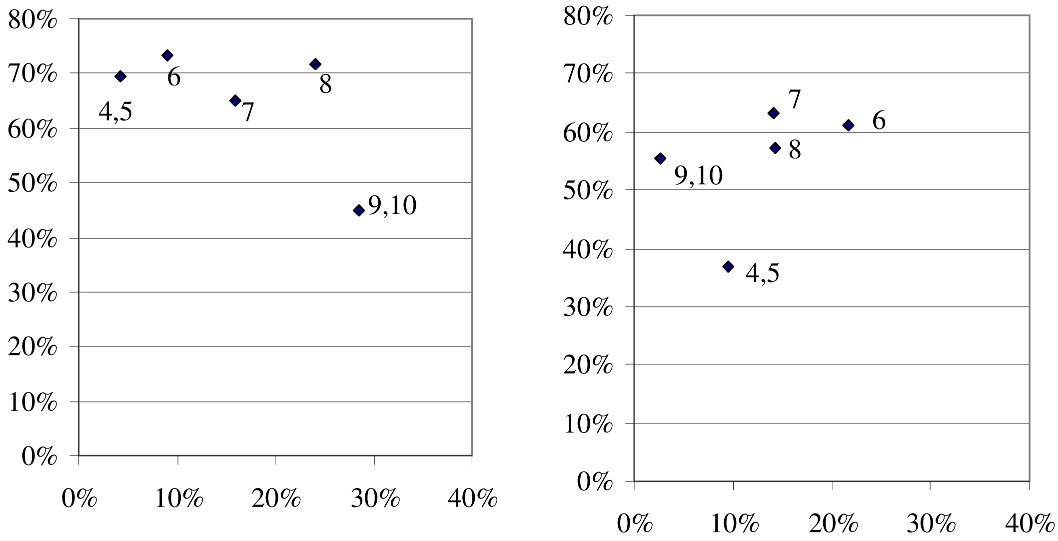
Steinle and Stacey (2003b) provided alternative figures for the student-focussed prevalence of A and A1, in an attempt to overcome the project effect. By noting the increasing test-focussed prevalence of both A and A1 in each grade on first tests, they argued that the test-focussed prevalence of A and A1 in Grade 6 and Grade 10 were better indicators of the student-focussed prevalence in the primary school and secondary school, respectively. Hence, the best estimates (provided in brackets in Table 9.3) for the student-focussed prevalence of A and A1 in the primary school are 40% and 30% respectively, and the corresponding figures for the secondary school are 80% and 70% respectively.

Figure 9.4 provides the condensed test histories of A1, A2, A3 as discussed in section 9.1 (i.e. the distribution of the second and third non-U tests, given that the first non-U test is A1, A2 or A3). The similar feature of these three graphs is the tall column indicating that both the second and third non-U tests are A; the differences between these graphs are discussed below. Figure 9.5 provides the persistence-proximity maps for both A2 and A3 for various grades and these will also be discussed below.



a) A1 (n=733) b) A2 (n=86) c) A3 (n=38)

Figure 9.4: Distribution of second (left-hand axis) and third (right-hand axis) tests in condensed test histories, given first test is A1, A2 or A3



a) A2 b) A3

Figure 9.5: Persistence-proximity maps for A2 and A3 by grade

9.3.1 The fine code A1

In this thesis we have referred to a student who makes very few errors on the Decimal Comparison Test (i.e. the test is allocated the code A1) as a *task expert*, as a reminder not to make assumptions about the student's ability to complete other tasks involving decimals. Furthermore, no claim is made on the reasoning or algorithm that guided their choices on any particular item. Successful completion of the DCT is a *necessary* but not *sufficient* condition for overall expertise with decimal notation. Hence all of the estimates of expertise, from any version of the DCT, are overestimates of the prevalence of students with a thorough understanding of decimal notation. Note that Figure 9.3 indicates about 70% of students in Grade 10 are *task experts*.

The phenomenon of regression (which affected about 1 in 5 eligible students in this study) confirms that not all students completing a test coded as A1 are able to do this in all subsequent tests. In particular, the high rates of regression to both A2 and A3 confirm that some of the A1 students are using an incomplete algorithm described in Chapter 8; such as *truncating or rounding to one or two decimal places*, or *partial left-to-right digit comparison* (where the student doesn't realise that the space at the end of 17.35 can be replaced by a zero to complete the comparison with 17.353).

This result was confirmed by considering the order that different codes appear within students' test histories in Chapter 7. Of the 121 students with both A1 and A2, about 40% have their first A2 code *after* their first A1; similarly for the 90 students with both A1 and A3 tests.

Hence, students who have an impoverished conceptual understanding inflate the prevalence of A1 in Figure 9.3 and Table 9.3; they appear as experts on one test (by following an algorithm to assist them compare decimals) but then are unable to repeat this on a later test.

9.3.2 The fine codes A2 and A3

In section 3.2.3, the term *money thinking* was used to refer to students who were unable to complete comparisons between decimal numbers where the digits in the first two columns after the decimal point were identical (e.g. 17.353 / 17.35). In Chapter 8, five incomplete algorithms were discussed which lead to tests being coded as A, but not necessarily A1. In other words, students completing tests coded as A2 or A3 may have a particular way of thinking, (such as *money thinking*), or may be following incomplete algorithms (such as round to two decimal places). It is not possible to determine in this thesis which students are using an analogy (such as the context of money) and which students are using an algorithm. Indeed, it may be the case that the particular context (e.g. money) has supported the use of the algorithm.

What is clear, however, is that the large proportion of students in secondary school involved in the codes A2 or A3 (26%) indicates teaching that either provides limited contexts or focusses on rules (rather than an understanding of place value).

In particular, there is evidence to support the hypothesis that A2 students are harbouring a *latent* misconception (S behaviour) to deal with the failure of their incomplete algorithm. In contrast with the A1 and A3 students, for example, Figure 9.4b shows that about 10% of the students with an A2 on their first non-U test have an S test on *both* their second and third non-U tests (see the column at the front). Furthermore, it will be noted later that older S1 students have an increasing tendency to move to A2 on their next test.

While Figure 5.6a indicated that the rate to S (on the next test) for A2 students was higher than the rate for both A1 and A3 students, Figure 9.6 provides a more detailed analysis by considering the grade of the current test. The increased rate of A2 to S in Grade 7 (especially in comparison with A1 and A3) supports the hypothesis of interference of new teaching.

It is hypothesized that students completing tests allocated the code A3 may also be using an incomplete algorithm to compare decimals and in this case, it is L behaviour. Evidence to support this claim can be found in Table 9.2. Of the 239 students with at least one A3 test (who have more than one test in total) nearly 30% tested as L1 at some time in the study, which is nearly twice the level for A1 and A2 students.

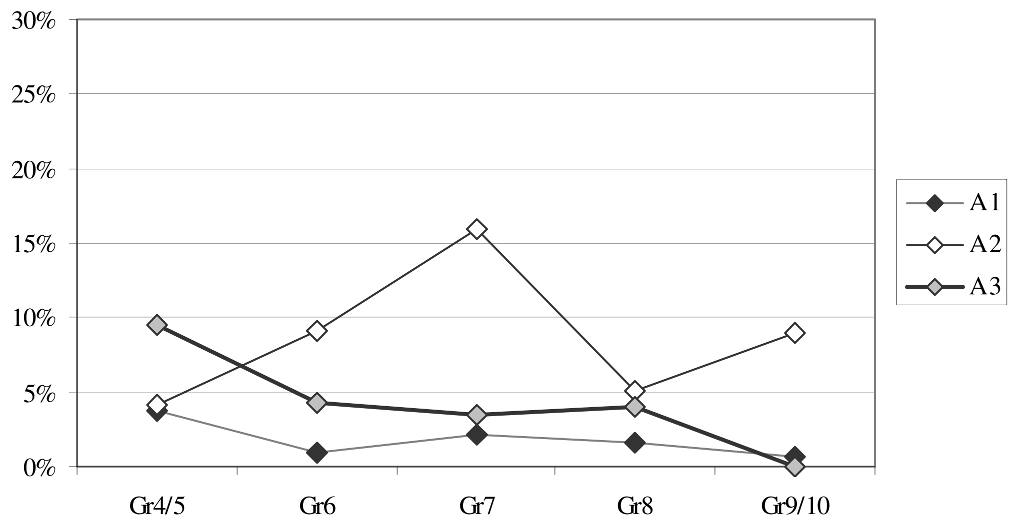


Figure 9.6: Rate to S on next test from A1, A2 and A3 by grade of current test

How *advanced* is the thinking behind the tests allocated to A2 or A3? Figure 9.2 shows that A2 and A3 are closer to A1 than most other codes, however, Figure 9.5a shows that persistence in A2 increases steadily with grade. Hence, it appears that these students are becoming more comfortable with their current way of thinking and/or algorithms being used and some of these students are likely to leave school without realising that there is something that they do not know. For example the procedure of rounding decimals to two decimal places may well be reinforcing the idea that longer decimal numbers are some sort of error and the only appropriate action to remove these errors is to round to two decimal places.

Another piece of evidence that A2 is not advanced is obtained by a comparison of A2 and S3. Table 5.14 indicates that there were 40 students with both A2 and S3 in their test histories. Table 7.4 shows that over 40% of these students had their first A2 *before* their first S3, (which is not expected if A2 is more advanced than S3). In particular, S3 was found to be one of the most primitive codes, by considering its proximity to A1 (see Figure 9.2). Hence, we believe that the A2 code is not particularly advanced; it is possible to do well on the DCT by only considering the first one or two decimal places.

9.4 The L cluster of codes

Figure 9.7 provides the improved test-focussed prevalence of L1, L2 and L4 by grade and clearly shows that L1 dominates the coarse code L and that the prevalence of L tests decreases with grade. At each grade level the percentage of L students is approximately 60% of the percentage at the previous grade level. The prevalence of L in Grade 8 is 15%, which, coincidentally, is the exact proportion of the Australian sample in the TIMSS-R study which chose 0.5 in Table 2.3.

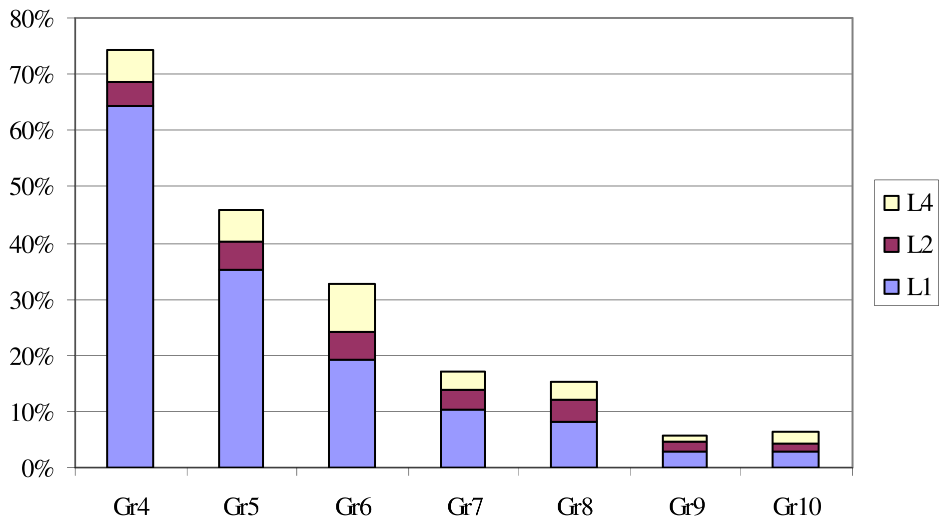


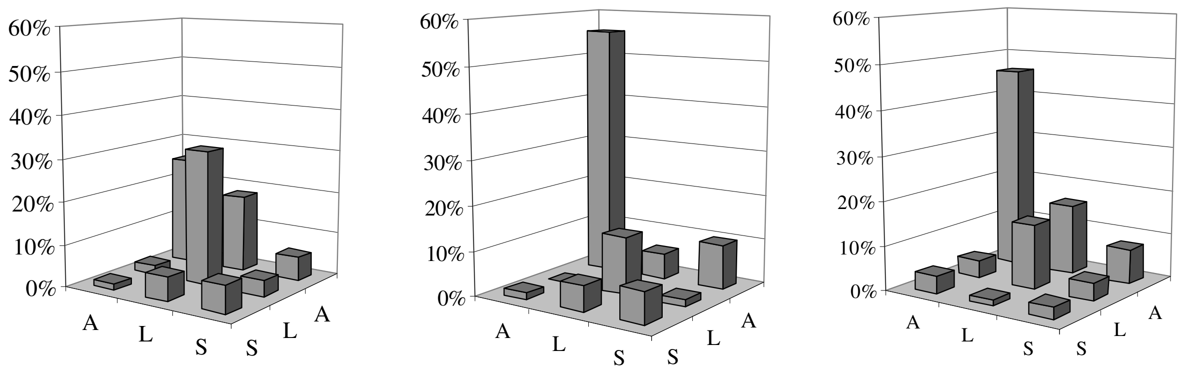
Figure 9.7: Improved TFP of L1, L2 and L4 by grade

Table 9.4 provides the improved student-focussed prevalence of these codes for each school level and reveals that about 1 in 5 secondary students test as L at some time. Hence, the code L is not completely confined to just primary schools. In general, students either stay or leave, but don’t tend to return once they leave L.

Table 9.4: Improved SFP of L, L1, L2 and L4 by school level

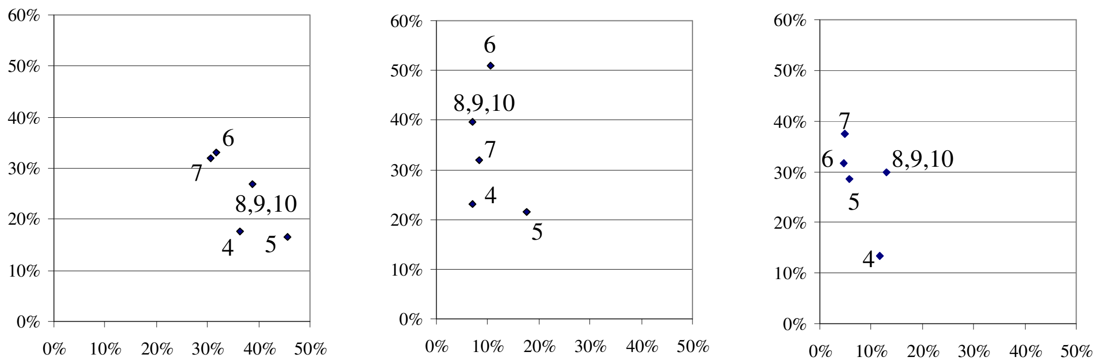
School Level		L	L1	L2	L4
Primary	(n=333)	71	62	19	17
Secondary	(n=682)	21	12	7	5

Figure 9.8 provides the distribution of next two non-U tests, given first non-U test is L1, L2 or L4 and Figure 9.9 provides the map of persistence against proximity to A1 for these codes. Differences between these graphs will be discussed below.



a) L1 (n=375)　　b) L2 (n=70)　　c) L4 (n=76)

Figure 9.8: Distribution of second (left-hand axis) and third (right-hand axis) tests in condensed test histories, given first test is L1, L2 or L4



a) L1　　b) L2　　c) L4

Figure 9.9: Persistence-proximity maps for L1, L2 and L4 by grade

9.4.1 The fine code L1

Students using *whole number thinking* will be allocated the code L1 on the DCT. These students treat a decimal number as a pair of whole numbers and so choose 4.9 as smaller than 4.11. There are two variations within this group (*numerator focussed thinking* and *string length thinking*), but these have not been separated within this thesis.

Whole number thinking was confirmed to be the most primitive way of thinking, (in particular in the primary school) with the lowest rate to expertise on the next test; see Figure 9.2. Furthermore, L1 was shown to be more than twice as likely to occur *before* another code in a student's test history, than *after* it (see Table 7.4).

Many students find it very hard to break free from this way of thinking and test as L1 repeatedly. Figure 9.9a indicates that persistence in L1 is over 30% for students of all grades. The fact that such students make few errors when dealing with equal length decimals may be the reason for this persistence, as many decimal activities in classrooms are embedded in a context of measurement where the decimals are of equal length. Students not making errors will not realise that they have something more to learn. Figure 9.8a shows clearly that the most likely scenario for the two (non-U) tests following an L1 test is that they are both in L.

9.4.2 The fine code L2

As explained in section 3.2.1, students using either *zero makes small thinking* or *column overflow thinking* are expected to complete tests coded as L2. A student using *zero makes small thinking* is basically using *whole number thinking* with the addition of an isolated fact, (i.e. that a zero after the decimal point makes the number small). A student with *column overflow thinking* incorrectly generalises the relationship between the number of zeros after the decimal point and the column names. With numbers such as 0.3, 0.04 and 0.005, the number of zeros after the point *can be used* to infer the size of the pieces (tenths, hundredths and thousandths, respectively). This relationship *does not hold*, however, for multi-digit numbers; as 0.37 is *not* 37 tenths, 0.047 is *not* 47 hundredths and 0.0057 is *not* 57 thousandths. A student with *column overflow thinking* is generalising from correct decimal place value knowledge and hence is more advanced than student with *zero makes small thinking*.

No evidence has yet been provided in this thesis to either support or reject the hypothesis that the code L2 is being allocated to two groups of students with different ways of thinking. Consideration of the varying rates to A1 on the next test, by current grade, may now provide support for this hypothesis. Figure 9.9b indicates that the proximity to A1 for the youngest L2 students in Grades 4 and 5 is approximately 20% (based on a combined sample of over 50 students), compared with 30% to 50% of older students. This could indicate that, compared with the older students, more of the younger L2 students are using the less sophisticated *zero makes small thinking.*

In addition, the increased proximity to A1 for the L2 students in Grade 6 (over 50%) is based on a reasonably sized sample (51 students). This figure is higher than for any of the other L codes in any grade, and, as we shall see in the next section, is also higher than any of the S codes in any grade (with the exception of S1 students in Grade 6). This indicates that the teaching that occurs in Grade 6 is more likely to help the students using these ways of thinking (*column overflow thinking* and *zero makes small thinking)* than teaching in any other grade and for any other ways of thinking behind the L and S behaviours (with one exception).

Figure 9.8b shows that the most likely scenario for the two tests following an L2 test is AA; this confirms the place of L2 above both L1 and L4 in the hierarchy (as was found in Chapter 7). In fact, *zero makes small thinking* was predicted to be more advanced than *whole number thinking* as it is based on *whole number thinking* with the addition of an isolated fact, (i.e. that a zero after the decimal point makes the number small). Similarly, from a theoretical consideration, *column overflow thinking* is more advanced than *whole number thinking.* The position of L2 as more advanced than L1 was confirmed when the test histories of students with both L2 and L1 were examined. Over 80% of these students had their first L1 code before their first L2 code; similarly with L4 and L2 (see Table 7.4).

9.4.3 The fine code L4

Students completing tests coded as L4 are likely to be in transition between L1 and L2. An examination of the test histories of students with both L4 and L1 indicates that the first occurrence of L4 is more than twice as likely to come *after* the code L1, as *before* it (see Table 7.4). For students with both L4 and L2, the reverse is true; that is, the first occurrence of L4 is more than twice as likely to come *before* the code L2, as *after* it. This confirms the place of L4 as between L1 and L2 in a hierarchy. Furthermore, Figure 9.8c indicates that the percentage of AA tests after L4 (46%) is between the lower value for L1 (26%) and the higher value for L2 (56%). Of the twelve students in this study whose test histories contained all three fine L codes, seven of these students had these tests in the order as predicted: L1 before L4 before L2 (see Table 5.11).

9.5 The S cluster of codes

In this study, the distribution of S1, S3 and S5 within the coarse code S was found to be about one quarter, one half and one quarter, respectively. Due to this flatter distribution of the fine codes within S, (compared with both A and L), the following discussion will concentrate more on the coarse code S.

Figure 9.10 provides the improved test-focussed prevalence of S1, S3 and S5 by grade. The highest prevalence of these S codes is in Grades 6, 7 and 8. The prevalence of S3 in Grade 8 is approximately 10% which is substantially less than the 22% of Grade 8 students in the Australian sample who chose 0.625 in the TIMSS-R item (see Table 2.3).

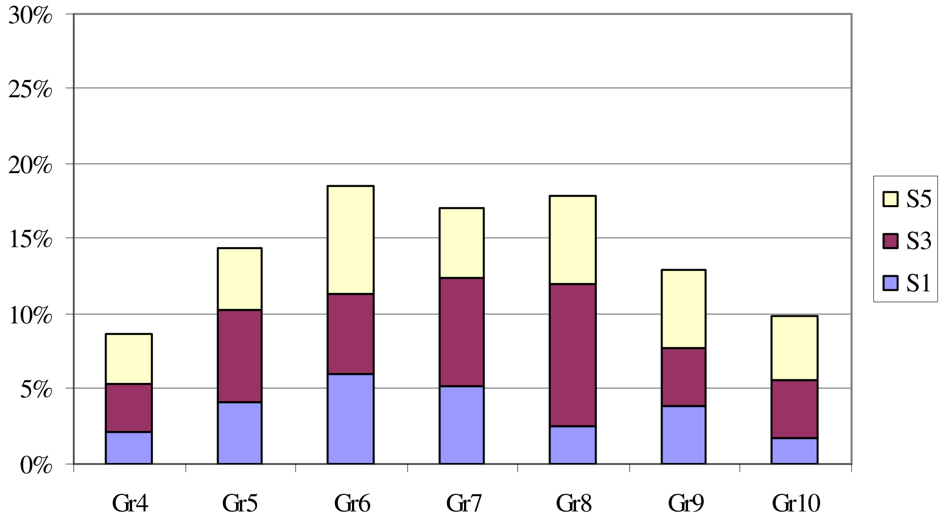


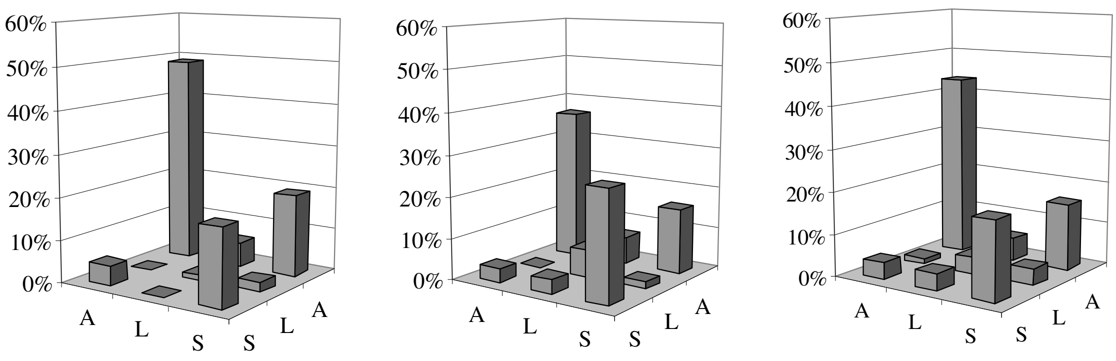
Figure 9.10: Improved TFP of S1, S3 and S5 by grade

Table 9.5 provides the improved student-focussed prevalence of these codes for each school level. Approximately 1 in 3 students will be involved in S tests during their years at primary school, and approximately 1 in 4 in secondary school.

Table 9.5: Improved SFP of S, S1, S3 and S5 by school level

School Level		S	S1	S3	S5
Primary	(n=333)	35	17	16	9
Secondary	(n=682)	28	10	17	10

Figure 9.11 provides the distribution of next two non-U tests, given first non-U test is S1, S3 or S5. The three most likely combinations in these three graphs are: AA (back corner), SS (front corner) and SA (right hand corner); the differences between these graphs are discussed later. Figure 9.12 provides the persistence-proximity map for S1, S3 and S5 and is also discussed later.



a) S1 (n=88) b) S3 (n=121) c) S5 (n=76)

Figure 9.11: Distribution of second (left-hand axis) and third (right-hand axis) tests in condensed test histories, given first test is S1, S3 or S5

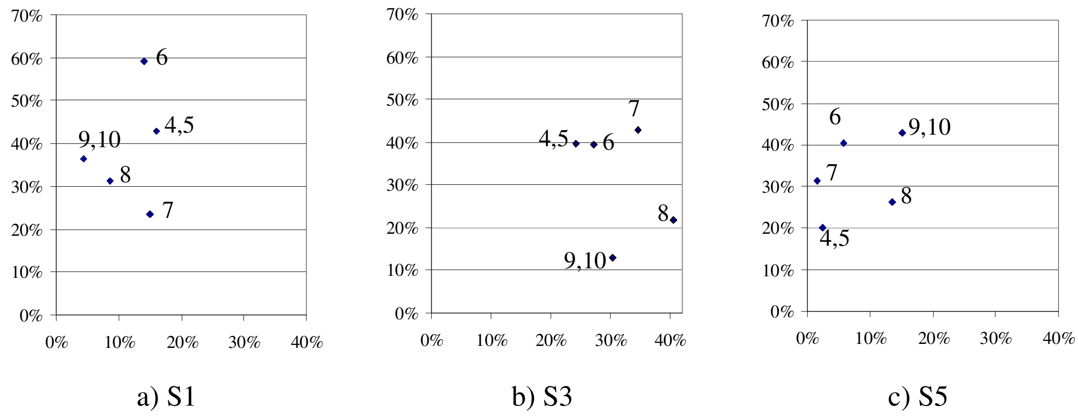


Figure 9.12: Persistence-proximity maps for S1, S3 and S5 by grade

Figure 9.13 provides the rate to S on the next test by students who are currently in any of the non-A1 codes. (This graph is a result of combining the rates to S1, S3 and S5 from the non-A1 codes in various tables in Appendix 7.) As mentioned in Chapters 2 and 3, it is proposed that interference from new teaching of both negative numbers and scientific notation in Grades 7 and 8 is the most likely reason for the higher rates to S.

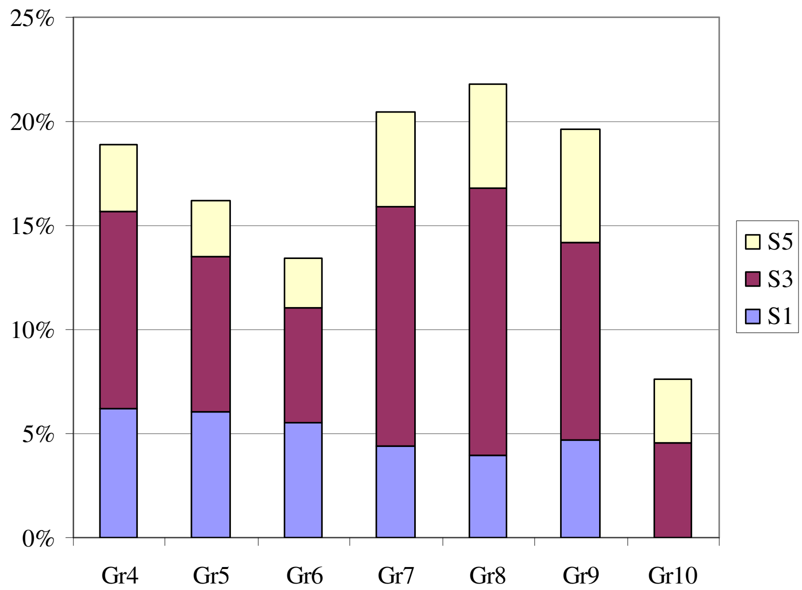


Figure 9.13: Rate to S1, S3 and S5 on next test from non-A1 test by current grade

9.5.1 The fine code S1

Students using either *denominator focussed thinking* or *place value number line thinking* are expected to complete tests that are allocated the code S1. These students typically choose the shorter decimal as the larger, but can order equal length decimals correctly.

A student with *denominator focussed thinking* knows that column names to the right of the decimal point indicate smaller and smaller parts. They incorrectly generalise, however, from 1 tenth is larger than 1 hundredth, to any number of tenths is larger than any number of hundredths. So, they believe that 4.82 < 4.3 as hundredths are smaller than tenths. A student with *place value number line thinking* has the same responses on the DCT but some sort of image of a number line may be appearing to these students. Figure 3.2 is an attempt to display how the place value columns are morphed into an incorrect number line.

While the relative heights of the AA columns in Figure 9.11 suggest that the code S1 is the most advanced S code, it was found in Chapter 7 (and can also be seen in Figure 9.2b) that the rate to A1 from an S1 test in secondary school was so low that S1 was ranked lower than L1 on this measure in the secondary school.

Figure 9.12a indicates that 60% of the S1 students in Grade 6 who change codes (based on a sample of 49 students) move to A1 on the next test. This is higher than any of the other S codes in any grade, and higher than the any of the L codes in any grade; L2 in Grade 6 is the closest with 50%. This indicates that the teaching that occurs in Grade 6 is more likely to help the students using these ways of thinking (*denominator focussed thinking* and *place value number line thinking*) than teaching in any other grade and for any other ways of thinking behind the L and S behaviours.

Where do S1 students go if not to A1? These students have an increased tendency (with grade) to move to A2, S3 and S5, (see Table 26 in Appendix 7). As noted earlier, this provides additional evidence that A2 students are using algorithms to compare decimals, but harbour an S behaviour.

9.5.2 The fine code S3

Students using either *reciprocal thinking* or *negative thinking* are expected to complete tests that are allocated the code S3. These students typically choose the shorter decimal as the larger, and cannot order equal length decimals correctly. Both of these ways of thinking are expected to be rather primitive, as like *whole number thinking*, they treat the decimal portion as a second whole number. The difference is that both students with *reciprocal thinking* and *negative thinking* then choose the smaller whole number in the decimal portion to give the larger number. For example, given 0.3 and 0.4, they would choose 0.3 as larger by some analogy with one third and one quarter, or negative three and negative four.

It was predicted that students who choose incorrectly on equal length decimals would make many errors in a classroom and hence be unlikely to persist with such thinking. This was shown not to be the case in Chapter 6. For example, Figure 9.12b shows that, persistence in S3 is high, 20% to 40% for all grades.

Grades 5 and 8 have slightly higher prevalence of S3 but inspection of the various school groups shows that some have quite elevated levels of S3 (about 10%) in one or both of Grades 5 and 8. We propose that the S3 in Grade 5 is *reciprocal thinking* and that the students are using their fraction understanding to try to understand the newly introduced decimal notation. In contrast, the reason for the presence of S3 in junior secondary school is proposed to be due to *negative thinking*.

9.5.3 The fine code S5

S5 is different to both A3 and L4 (the other *fine unclassified* codes) because of the larger proportion that it occupies within the coarse code; about 1 in 4 tests coded S are S5, compared with A3 (1 in 20 tests coded as A) and L4 (1 in 6 tests coded as L). It is proposed that students with one of the four ways of thinking that lead to S1 and S3, but who choose inconsistently on equal length decimals, are coded as S5.

Note that another alternative to explain S5 is a student who answers consistently on Types 5 and 6 (and hence would be coded as S1 or S3) but does not follow the predictions for the Type 3 items. It had been predicted that all S students would choose correctly on the Type 3 items (such as 4.7 / 4.08) as the shorter number is the correct answer, but Stacey and Flynn (2003) found a group of students who ignored zeros before selecting the shorter (and since they are interpreted as whole numbers, smaller) number as larger. They would therefore believe 4.7 is larger than 4.08 (ignore the zero, compare 7 and 8, choose smaller as larger). There are therefore likely to be variations of S thinking that could be traced in a more refined decimal comparison test.

9.6 The U cluster of codes

In this study, the distribution of U1 and U2 within the coarse code U was found to be 93% and 7%, respectively, hence clearly showing the dominance of U1 within the coarse U code.

Figure 9.14 provides the improved test-focussed prevalence of U1 and U2 by grade, Table 9.6 provides the improved student-focussed prevalence of these codes for each school level and Figure 9.15 provides the persistence-proximity maps for U1 (by grade) and U2 (not by grade).

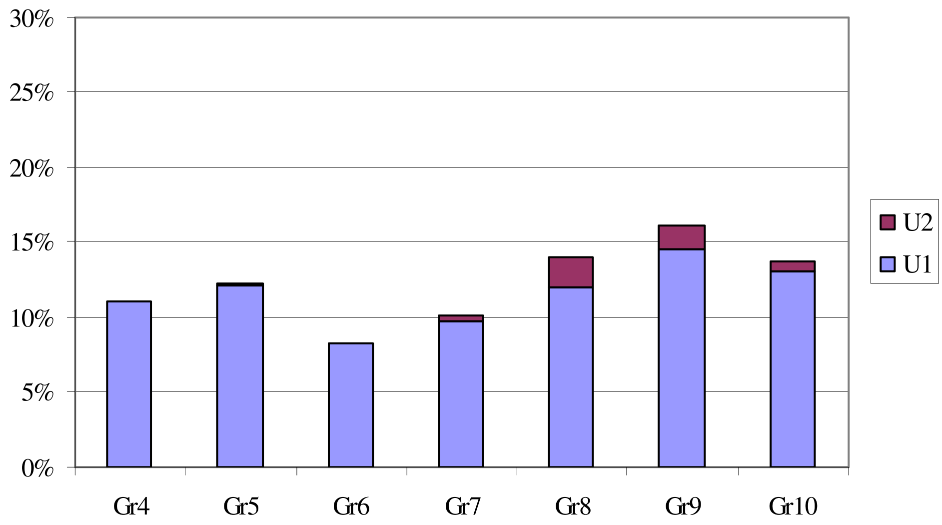


Figure 9.14: Improved TFP of U1 and U2 by grade

Table 9.6: Improved SFP of U, U1 and U2 by school level

School Level		U	U1	U2
Primary	(n=333)	44	44	1
Secondary	(n=682)	28	28	4

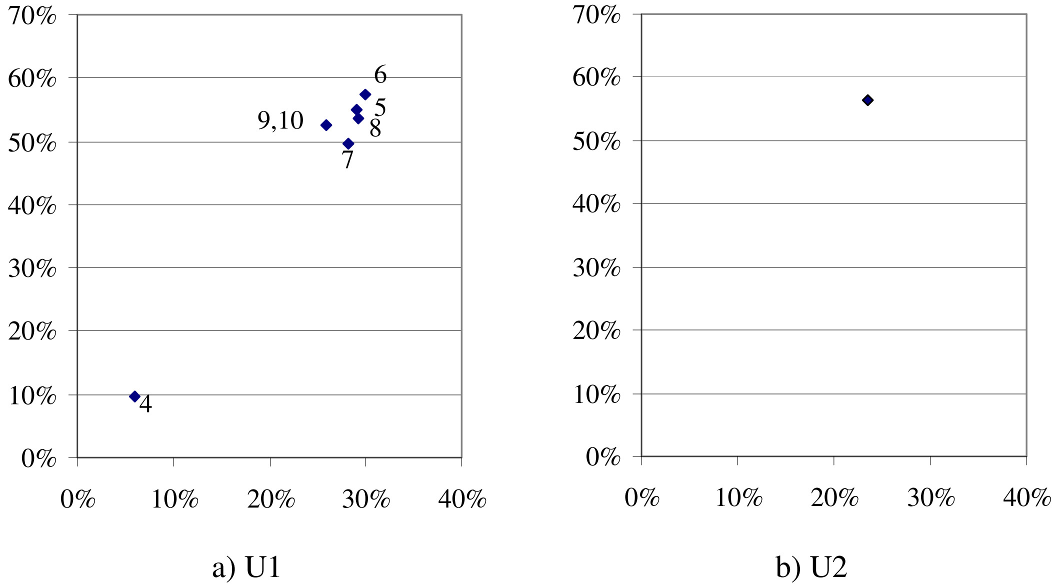


Figure 9.15: Persistence-proximity maps for U1 by grade and U2 (not by grade)

9.6.1 The fine code U1

Recall that U1 is not a way of thinking, rather a collection of tests that do not fit any of the predicted patterns of errors. Table 6.9 indicates that a higher proportion of students have scattered occurrences of U1 than for any other code; hence it is likely that such students are in transition between various other codes. Figure 9.2 indicates that U1 is more advanced than any of the fine L or fine S codes. Hence, for a student who has not yet reached expertise, being in transition is the optimal place to be.

The closeness of the points in Figure 9.15a is worthy of mention. Only U1 tests in Grade 4 have values of persistence (n=33) and proximity (n=31), which differ from the rest. This might be due to some Grade 4 students who had not yet been taught decimals; these students might have guessed randomly.

9.6.2 The fine code U2

Three possibilities have been proposed for the reason that students complete tests that are coded as U2. Firstly, an expert who misreads the instructions may choose the smaller rather than larger number throughout the test; secondly, an expert who is mischievous may decide to answer wrongly for fun; and thirdly, a student may be using a mis-rule. If they are aware of one of the algorithms for comparing decimals described in Chapter 8 (either annexing zeros to the shorter number and then comparing as whole numbers or using the partial left-to-right digit comparison), but have no conceptual basis for the reasoning behind the algorithm, then they may forget or invert the last step of the algorithm. In particular, after annexing zeros they may choose the smaller whole number in the decimal portion as representing the larger number, or after comparing digits from left to right they may then choose the smaller digit as indicating the larger number.

It has not been possible to distinguish between the three various scenarios presented above, but the over-representation of the code U2 (with both A2 and A3) in regression lends support to the mis-rule possibility. Steinle and Stacey (2002) proposed that this was likely after noting that older U2 students are less likely to move to A1 and more likely to stay U2. Furthermore, 1 in 6 of the U2 students move to U1 on their next test, which is not expected if the students are experts (who misread the instructions or are being mischievous).

9.7 Conclusion

Several new analyses were introduced in this chapter. Firstly, an analysis of students' condensed test histories was presented using 3-dimensional column graphs to illustrate the different paths by students over their first three non-U tests. Secondly, the new measure of association was introduced, which indicated the pairs of codes that were more likely to appear within particular students' test histories. Thirdly, persistence-proximity maps (2-dimensional visualisations of two measures from Chapters 6 and 7) were introduced. These analyses and visualisations aided the discussions within this chapter, which focussed on the ways of thinking behind the codes.

Improved estimates were obtained for both the test-focussed prevalence and the student-focussed prevalence of the coarse and fine codes. These estimates were obtained by using restricted samples; for example due to the test effect, only students' first tests were considered in the determination of the improved test-focussed prevalence, and only students who completed 4 or more tests in either primary school or secondary school contributed to the calculation of the improved student-focussed prevalence.

One of the main findings from Chapter 8 is that students use various algorithms to compare decimal numbers. What was confirmed in this chapter is that these students harbour either L or S behaviours, which are temporarily suspended by the use of an algorithm for comparing decimals. They then revert to these behaviours when their algorithm fails. In other words, they are receiving teaching that *covers over*, rather than *overcomes*, their misconceptions. Furthermore, the codes A2 and A3 (i.e. coarse A but not *task expert*, A1) are not sophisticated thinking. It is possible to answer many decimal comparison items correctly by comparing the tenths and hundredths places only.

Also confirmed in this chapter, is that the interference of new teaching in Grades 7 and 8 leads to an increased movement, among students who are not experts, to S behaviour on the next test.

Another point worthy of mention is that two particular codes (L2 and S1) have much higher proximity to expertise in Grade 6, compared with all other grades and all other fine L and fine S codes. There is something special about the particular combination of these ways of thinking and the teaching that occurs in Grade 6. Further research is required to investigate the reason for this anomaly.

CHAPTER 10 CONCLUSION

10.1 Summary of findings

One of the aims of this study was to provide improved estimates of the prevalence of the various misconceptions in students from Grades 4 to 10. This has been done in some detail in Chapter 9, and is only described here in general terms.

Many students with misconceptions about decimal notation judge the size of decimals by the relative number of digits that they contain. These students are said to be exhibiting L behaviour, if they consider *longer* decimals (i.e. more digits) to be larger, or exhibiting S behaviour if they consider *shorter* decimals (i.e. fewer digits) to be larger. By far the majority of students, when first encountering decimal numbers in Grades 4 and 5, exhibit L behaviour. As students move into older grades, fewer and fewer are left with L behaviour although about 1 in 5 secondary school students exhibit such behaviour at some time in their years in secondary school.

The reasons for S behaviour (discussed fully in Chapter 3) are more complex than for L behaviour; different students move in and out of S behaviour at different times. About 1 in 3 students will exhibit S behaviour at some time while they are in primary school and about 1 in 4 students while they are in secondary school.

This study has found that some students with S behaviour exhibit unexpected tendencies. They become more likely to persist with this behaviour in older grades and they become less likely to move to expertise. We believe that this is due to interference of new teaching, in particular, the introduction of negative numbers and scientific notation in junior secondary school.

In general, the misconceptions that are associated with interpreting the decimal portion as a whole number (referred to as *whole number thinking*, *reciprocal thinking* and *negative thinking*) are the hardest for students to leave. Of the students who completed tests that indicated such thinking, about one in six completed tests in a similar way more than two years later.

One of the findings of this study is that, among the students who made some errors on the Decimal Comparison Test, the students who were *most likely* to move to expertise on their next test were those who were *not exhibiting* L or S behaviours. Such students are possibly using a combination of different ideas, and may feel confused; appreciating that they have something to learn is the best state of mind for a non-expert. Such students may be more receptive to teaching than those with strongly held beliefs. The best estimate for the prevalence of expertise is 30% of students in Grade 6 and 70% in Grade 10.

The phenomenon of regression (students completing one test as an expert but unable to do so on a later test) was investigated. Of the sample of suitable students, one in five was involved in this phenomenon. The over-representation of certain codes in regression confirms that there are considerable numbers of students who are using particular (incomplete) algorithms to compare decimal numbers. Items such as the comparison of 0.45 and 0.453 (where one number is a truncation of the other) are the most likely to detect students who are using an incomplete algorithm. In this example, students using the *partial left-to-right digit comparison* algorithm will run out of digits to compare (as they do not know that the space at the end of the shorter number can be replaced by a zero). Similarly, students using the *round to two decimal places* algorithm will not be able to order these numbers. When an algorithm fails to provide a definite solution the students often resort to a *latent* misconception such as choosing on the relative number of digits (i.e. L or S behaviours). This confirms that rather than students moving away from these misconceptions, they are retained and are often hidden behind algorithms and procedures, (which, to an observer, indicate understanding). In other words, some students are receiving teaching that is *covering over* rather than *overcoming* misconceptions. In fact, uncovering the extent of this problem has been a major contribution of the thesis to the decimal misconceptions literature.

10.2 Evaluation of methodology

There are clear limitations to the test used to collect data for this study; the structure of the test and the particular items that it contains.

Firstly, the structure of the test; it is composed of a set of pair-wise comparison items, and does not include items which assess students' ability to perform operations with decimal numbers. Neither does it present contextualised problems or other probing tasks, such as insert a number between two given numbers. We know that it is possible to complete the test without an understanding of place value, just by following a rule; so all measures of expertise are over-estimates in some sense. The strength of the DCT is in identifying erroneous thinking and it is surprising that such a simple test can do this so effectively.

Intriguingly, students with a particular misconception (*reverse thinking*) seem to be frequently found in interviews but were found rarely with this test. We explain this by proposing that the format of the test is not conducive to making this error. In other words, there are features of a context that lead students to make certain errors.

Secondly, the items within the test; a subsequent version of this test (DCT0) has included items such as 0.6 / 0 which have generated quite unexpected responses. Stacey, Helme, Steinle, Baturo, Irwin and Bana (2001) found that 13% of a sample of over 500 pre-service teachers from four Australian universities chose 0 as larger than 0.6. Many of these students only made errors on the three items involving a comparison with zero; in other words, they would have been classified as experts on earlier versions of the test (including the DCT2 used in this study). Hence, students who believe that 0.6 is less than 0 inflate the prevalence of expertise found in this study. Recommendations for improvements to DCT2 are provided in section 10.4.

Despite the limitations of the structure of the test and the particular items that were included in this version of the test (DCT2), we have learnt much from students' responses to this test and the longitudinal data that has been presented in this thesis. Furthermore, not only do most students answer consistently on the items within the tests, this thesis has shown that they tend to answer consistently from one test to another. This confirms the validity of this test. The simplicity of this test and its power to diagnose misconceptions has produced a very positive reaction from teachers who have used this test in their classroom.

The 12 schools in this study were not chosen randomly but were selected to represent a mix of geographical and socio-economic areas of the city of Melbourne. Because of the commitment required to be involved in the project over several years, volunteer schools were used. While variation in the prevalence of expertise was closely aligned to socio-economic status for two schools in particular, we believe that the general findings from this study represent students from the wider community within Australia and possibly even internationally. In a separate study, a version of this test has been used with Japanese school students who exhibited similar misconceptions, although with variations in the prevalence (see Steinle and Stacey, 2001).

Some of the analyses (e.g. the prevalences) are affected by the nature of the sampling. The general agreement of the prevalences with the results of the relevant TIMSS-R item which were obtained on a properly constituted random sample (see section 5.2.1) give confidence in these prevalences, but it was not an aim of this thesis to obtain results which accurately represented the Australian population. The later improved estimates of prevalence account for some aspects of the sampling. Also the variation in school groups given throughout the thesis gives an indication of the range of school results and this can be more useful than averages. Analysis by class has not been done here; analyses show effects of teaching in small studies (e.g. Helme and Stacey, 2000a and 2000b).

On the other hand, many of the results in this thesis have been based on sub-samples of students who have an identified way of thinking about decimals and these results are less affected by the initial sampling. If we assume that their misconception is a major determinant of their subsequent behaviour and that their teaching has been representative of the teaching in Melbourne classrooms, then the composition of the initial sample of schools is of minor importance to the results.

The large sample (over three thousand students completed nearly ten thousand tests) enabled us to trace small effects and the student-focussed data showed that some of these are not as small as might first be thought.

One of the questions to be resolved throughout this thesis was the relative usefulness of the coarse and fine classifications of student misconceptions, (with 4 and 11 groups, respectively). Certainly the coarse codes are easier to use and explain to teachers, as the allocation of A, L, S and U depends on the pattern of responses to just two types of decimal comparison items (referred to as Type 1 and Type 2 items). Feedback from teachers suggests that the detection of students who exhibit L or S behaviour is sufficient for them to attempt to remedy the situation. On the other hand, within the students who are allocated the coarse code A, there are various groups of students who are using incomplete algorithms, as referred to above. To detect such students, additional types of items are required; in particular, Type 4 items involve the comparison of two numbers that have the same digits in the first two decimal places (such as 0.45 and 0.453).

In this thesis, a compromise between the coarse and fine codes was used in some analyses. Using a hybrid of coarse and fine classifications provides an intermediate amount of information; for example, considering pairs of codes using the coarse classification results in 16 possibilities (4x4), while using the fine classification results in 121 (11x11), and a hybrid results in only 44 (4x11).

Similarly, there has been a need to create the school level (rather than grade level) focus for the longitudinal measure of student-focussed prevalence of the misconceptions. This measure seems to be very successful as it tracks students over several years to give meaningful numbers as well using constructs that were already meaningful to teachers.

Other useful measures and definitions have been created that other researchers could use. For example, two measures were created in the investigation of hierarchy, (developmental ordering and proximity to expertise), and various measures of persistence were also defined. Issues that were due to sampling within this study were overcome to produce improved measures of prevalence.

No other study of the developmental patterns of misconceptions has been undertaken.

10.3 Implications for researchers

The prevalence of the various misconceptions as well as expertise (presented in Chapter 9) will provide a measuring stick against which further studies of teaching interventions can be compared.

It has been shown in this thesis that "zero is a very hard number". For example students using a *partial left-to-right digit comparison* algorithm do not know how to compare 3.64 with 3.641 as they do not know that the space after the 4 in the first number can be replaced by a zero to compare with the 1 in the second number. Further research should be undertaken to reveal the complete range of difficulties that students have with both the number zero as well as the digit zero.

Although the construction of the test is not part of the work done for this thesis, the thesis has many recommendations about improving it.

1. Pair-wise comparison is both useful and avoids overloading information processing capacity, so retain this in future versions.
2. The option of choosing equality on the test has been demonstrated by Steinle and Stacey (2001) as simple and effective. This allows additional types of items such as 0.8 / 0.80 and 3 / 3.0 (which are equal) as well as 0.7 / 0.07 (which some students consider to be equal).
3. The definitions of the various item types are provided in Appendix 1. Researchers wishing to construct other tests need to follow these definitions very carefully; what appears to be a very minor change can have a major impact on students' choices.
4. The items that involve a comparison of equal length numbers (e.g. 0.4 / 0.3) should be retained in future versions of the DCT. These were not used by earlier researchers but have been shown to effectively separate students with either *reciprocal thinking* or *negative thinking* from other students with S behaviour.
5. Some types of items were under-represented on DCT2. For example, there should be three items with different whole number portions (e.g. 1.3 / 0.86) to separate *decimal point ignored thinking* from *whole number thinking*. Furthermore, to separate *whole number thinking* into *numerator focussed thinking* and *string length thinking* would require three items such as 0.53 / 0.006.

6. Some types of items were not represented on DCT2. For example, Type 8 items such as 0.6 / 0 were used by Steinle and Stacey (2001) in a test referred to as DCT0 and were shown to be extremely useful in detecting the students who believe that as *all decimals are smaller than whole numbers*, then 0.6 is less than 0.
7. It would seem reasonable to drop *reverse thinking* from the classification, and hence remove the three Type 6 items (e.g. 0.42 / 0.35) to make room for other items.
8. Steinle and Stacey (2003a) identified two factors (number of decimal places exceeding two, and whole number portion being zero) which had a secondary but measurable effect on students' choices. These two factors interact to give four subtypes of items, α, β, γ and δ (see Appendix 1). Sets of more homogeneous items can now be constructed, by further refining Type 1 and 2 items, which should lead to better diagnosis.
9. Extra items could be included to separate the various incomplete algorithms; for example 0.457 / 0.4 is answered correctly by a student using the *round to one decimal place* algorithm, but a student who *truncates to one decimal place* may choose incorrectly.

This longitudinal study used a diagnostic assessment instrument and has provided a rich source of data that has been analysed in this thesis. We recommend that other mathematics topics also receive this focussed attention to reveal students thinking over periods of several years.

10.4 Implications for teaching

In this thesis, it has been demonstrated that many students persist with misconceptions from one grade to another as they move through school. Hence, it is clear that the normal teaching that most students receive is inadequate to remove such misconceptions. Hiebert's (1987) comment is pertinent,

> Introducing decimal addition by saying "These are just like whole number problems after you line up the decimal points" is a quick way to get good performance *in the short run* but it is counterproductive *in the long run*. The findings from research encourage us to adopt a long-term view and take the time to develop meaning at the outset. (p22)

Teachers will need to examine critically the texts and activities that they use in their classrooms, to determine if students with various misconceptions are able to answer correctly and hence go undetected. Furthermore, as noted by Swan (1990), teachers need to reconsider their "implicit beliefs",

> Most common mathematical texts and teaching practices seem to be based on the implicit belief that repeated rehearsal of facts and skills somehow result in better conceptual understanding. (You understand decimals better if you learn to add them up successfully…) (p45)

Teachers (and textbook writers) need to emphasize the additive structure of our base ten numeration system; that is 4.37 = 4 ones + 3 tenths + 7 hundredths. Awareness of this structure makes a comparison of 4.3 and 4.37 almost trivial. Students who are encouraged to "add zeros" and then compare 30 with 37 are not receiving any lasting teaching; such shallow teaching is likely to result in only short-term gains. Students need to be assisted with the concept of re-unitising; 3 tenths is equal to 30 hundredths.

Indeed, discussions of 30 and 37 may reinforce the misconceptions that involve students treating the decimal portion of a number as a whole number. This thesis found that these misconceptions are the hardest for students to leave.

While using decimal numbers in a context may be considered as good teaching, a note of warning is given. The most common contexts are in measurement and are actually systems of units and subunits. For example, 64.37m (i.e. one decimal number, one unit) can also be thought of as 64 metres and 37 centimetres (two whole numbers, two units). Note how the one number (64.37) has been split into two separate whole numbers (64 and 37). In another context, 64.37kg might be thought of as 64 kilograms and 370 grams. Focussing on the digit 3 in these numbers; first it was 3 *tenths*, but has also been read as 3 *tens* (in thirty seven) and then 3 *hundreds* (in three hundred and seventy) in the two different contexts. Teachers need to be aware that the treatment of decimals in a measurement context does not necessarily teach about decimal numbers at all! Once the subunit has been identified in a given context, students revert to a system of two whole numbers, in which there is no need to partition into smaller and smaller amounts. Hence, students do not see the extension to further place value columns to the right or the fact that between any two (decimal) numbers there are an infinite number of decimals that can be written.

A useful non-measurement context to discuss decimals is the Dewey Decimal System for locating books in a library. A student looking for a book numbered 510.316 will need to know that is will be found *after* 510.31 and *before* 510.32. Hence, ordering of decimals can be explored in this context, but this decimal system does not have other measurement attributes; for example the books numbered 510.34, 510.35 and 510.36 are unlikely to be evenly spaced about the shelves.

Teachers need to be aware that always rounding the result of a calculation to two decimal places can reinforce the belief that decimals form a discrete system and that there are no numbers between 4.31 and 4.32, for example. Teachers of students in junior secondary school also need to be aware that the introduction of negative numbers can interfere with students thinking about decimals as it can reinforce the idea that the decimal portion of a number is a whole number. Furthermore, the introduction of scientific (or exponential) notation and its use with small numbers (such as $1.3\text{x}10^{-6}$) can reinforce the idea that decimals are linked to negative numbers.

To overcome some of the difficulties listed above, teachers (and text-book writers) need to use ragged decimals wherever possible (i.e. with varying numbers of digits after the decimal point) and emphasize the place value of the digits. The use of a number line is strongly recommended, as it is possible to incorporate the various sets of numbers that tend to be dealt with separately in the curriculum, for example, whole numbers, fractions, negative numbers and decimals. The number line appears to be one of the few models that are useful for discussing the density of decimals. The material that Steinle, Stacey and Chambers (2002) refer to as LAB (Linear Arithmetic Blocks) is extremely useful, as it can be used to build number lines, whilst also strongly embodying the base

10 structure. This is what makes it different from an ordinary number line, which has no evident base ten structure. This CD-ROM also contains many ideas and activities for teaching decimals.

This thesis found that, of the non-expert students, those who are most likely to move to expertise on the next test are those who do not hold onto a particular misconception. In other words, these students do not exhibit either L or S behaviour. Seminal work by Swan (1983) proves that providing students with teaching that generates "cognitive conflict" can result in better outcomes for the students than "positive only" teaching. Work conducted in parallel to this thesis has investigated the possibilities for using artificial intelligence to track students' misconceptions and then to provide situations that generate cognitive conflict for individual students; see, for example Stacey, Sonenberg, Nicholson, Boneh and Steinle (2003). The complexities of the misconceptions revealed through the work in this thesis and elsewhere lead us to conclude that it is difficult for teachers to perform a thorough diagnosis in "real time".

I would like to complete this thesis on a positive note. Understanding misconceptions is one important step to improving instruction. They exist in part because of students' overriding need to make their own sense of the instruction that they receive and they can be overcome by teaching that pays attention to them. As Graeber and Johnson (1991) commented,

> It is helpful for teachers to know that misconceptions and buggy errors do exist, that errors resulting from misconceptions or systematics errors do not signal recalcitrance, ignorance, or the inability to learn; how such errors and misconceptions and the faulty reasoning they frequently signal can be exposed; that simple telling does not eradicate students' misconceptions or "bugs" and that there are instructional techniques that seem promising in helping students overcome or control the influence of misconceptions and systematic errors. (p1-2)

GLOSSARY

Term (Chapter)	**Description**
annexe zero algorithm (Ch8)	The *annexe zero algorithm* refers to the procedure of writing zeros on the right end of the shorter decimal until both decimal portions have the same length, and then comparing the decimal portions as whole numbers. So, for example, 0.75 is seen to be less than 0.8 because 75 is less than 80.
association (Ch9)	The *association* of code X with code Y is the proportion of students with code X (and a total of two or more tests) who have code Y in their test histories.
behaviour (Ch2)	The two main incorrect *behaviours* exhibited by students, when asked to compare a set of decimal numbers, are: Longer-is-larger (L behaviour), choosing the decimal with the *most* digits after the decimal point as the largest, and Shorter-is-larger (S behaviour), choosing the decimal with the *fewest* digits after the decimal point as the largest.
codes (coarse, fine) (Ch3)	A *code* is allocated to each test on the basis of that student's performance on various item types. The 4 *coarse* codes (A, L, S and U) are intended to represent behaviours and the 11 *fine* codes (A1, A2, A3, L1, L2, L4, S1, S3, S5, U1 and U2) are intended to represent the underlying ways of thinking.
cohort (Ch4)	A *cohort* refers to a group of students who are in the same grade in the same calendar year and are expected to move through the grades together; for example, students in Grade 6 in 1997 are expected to be in Grade 7 in 1998. Cohorts are numbered according to the calendar year that they would have been expected to start school.
condensed test histories (Ch9)	The remaining (ordered) set of tests completed by each student after the U tests have been removed from the complete test histories. For example, a student with a test history of [L,U,S,U,A,A] has a condensed test history of [L,S,A,A].
consecutive tests (Ch5)	A student who has a test history of [T_1,T_2,T_3] is considered to have two pairs of consecutive tests: (T_1, T_2) and (T_2, T_3). Note that these may or may not have been completed in consecutive semesters.

decimal, *decimal number* *(Ch1)*	The terms *decimal number* or just *decimal* will be used in this thesis to indicate that a decimal point has been used to write the number, whether or not the number is greater or less than one, and whether it is rational (finite or infinite length) or irrational. Unless specifically mentioned in this thesis, all decimal numbers are of finite length.
decimal comparison items (Ch1)	The term *item* refers to a pair-wise comparison item (such as comparing 0.8 and 0.75)
decimal comparison tasks (Ch1)	While various researchers have used items that involve either ordering a set of decimals, or selecting the largest or smallest decimal from a set, in this thesis these will both be referred to as *decimal comparison tasks.*
Decimal Comparison Test (Ch3)	*Decimal Comparison Test* (DCT) is created by a sequence of decimal comparison items.
decimal portion (Ch1)	A number such as 64.530 is said to be composed of two portions: the *whole number portion* (64) and the *decimal portion* (530).
developmental ordering (Ch7)	By considering the typical ordering of two codes within students' test histories, we can create a *developmental ordering* of the codes. For example, in almost every case that a student has both L1 and A1 codes in their test history, the L1 code occurred before the A1 code. Hence, L1 is *developmentally younger* than A1, or A1 is *developmentally older* than L1.
errors (Ch3)	There are two types of *errors* that students make that we wish to distinguish between. A student with a misconception will make *systematic* (predictable) errors as they answer the DCT. On the other hand, some students make *careless* errors, which are not systematic.
expert algorithms (Ch1)	Two *expert algorithms*: Both algorithms first compare the whole number portions, but operate differently if the whole number portions are the same. See *annexe zero algorithm* and *left-to-right digit comparison algorithm.*

expertise (Ch8)	In Chapter 8, expertise is defined with respect to the fine codes; a student who completes a test which is allocated the code A1 is demonstrating expertise. In earlier chapters, (for example, section 7.4 Proximity to expertise), some analyses are undertaken with both the coarse codes and the fine codes and in this case, expertise is used loosely to refer to first A and then A1.
hierarchy (Ch7)	An ordering of the codes from the most *primitive* to the most *advanced*? It is proposed that any such ordering does *not* represent a series of steps or stages that a student must pass through on the way to expertise. Two analyses are used and compared: *developmental ordering* and *proximity to expertise.*
item types (Ch3)	Items are grouped into *item types*, when they elicit the same response (whether correct or incorrect) as various individuals complete the DCT. Defined in Appendix 1.
left-to-right digit comparison algorithm (Ch8)	The *left-to-right digit comparison algorithm* refers to the procedure of comparing digits in corresponding columns, moving from left to right until a larger digit is found. (So, 0.75 is seen to be less than 0.8 because 7 is less than 8.) The complete algorithm requires knowledge of what to do when no difference in the digits is found (i.e. when run out of digits to compare, the space in the shorter decimal can be replaced by a zero); without this knowledge, it is referred to as a *partial* algorithm.
length of a decimal number (Ch1)	The *length of a decimal number* is the number of digits explicitly displayed in the decimal portion, so the length of 64.370 is 3.
meta-data (Ch4)	Data about the data.
misconception (Ch1)	The use of the term misconception will be limited to introductory chapters in this thesis and used in a general way; the more precise terms *behaviour* and *way of thinking* will replace it in later chapters.
one-digit decimals, two-digit decimals, (Ch1)	Decimal numbers such as 0.3 and 1.3 are referred to as *one-digit decimals*, and 0.54, 0.80 and 3.84 are referred to *as two-digit decimals*, etc.
partial left-to-right digit comparison algorithm (Ch8)	This is an incomplete version of the *left-to-right digit comparison algorithm.*

persistence (Ch6)	*Persistence* refers to the tendency to retesting in the same code. *Persistence over one semester* is the proportion of students who retest in the same code on the next test, which is usually in the next semester (on average, 8 months later). *Persistence* can also be determined over longer periods.
persistence-proximity maps (Ch9)	Two dimensional graphs: *persistence* on X-axis and *proximity to expertise* on the Y-axis.
prevalence (Ch5)	See test-focussed prevalence and student-focussed prevalence.
project effect (Ch5)	The *project effect* is defined to be the improved performance on the DCT that is due to having completed the test before. It is the difference between the prevalence of expertise (A1) for two groups of students in the same semester (those on their first test and those on a *subsequent* test).
proximity to expertise (Ch7)	The *proximity to expertise* is defined as the rate of movement to expertise (A or A1) on the next test, given that the student does not persist in their current way of thinking.
ragged (Ch1)	The term *ragged* will be used to describe a set of decimals that are not all written to the same number of decimal places. For example, the set {0.8, 0.75} is a set of ragged decimals, while the set {0.80, 0.75} is not. Thus a set of decimals is described as either *equal length* or *ragged*.
regress (Ch8)	A student is said to *regress* if they have a test coded as A1 followed by a test with a different code. Hence, regression is detected when two consecutive tests by a student are examined.
regression pair of tests (Ch8)	The two consecutive tests indicating regression, i.e. (A1, non-A1) are referred to as the *regression pair of tests.*
regression test (Ch8)	The second test in the *regression pair of tests* is referred to as the *regression test.*
repair (Ch8)	A student who uses an incomplete procedure will *repair* the procedure when it reaches an impasse.
scattered occurrences *of codes (Ch6)*	A student may have two occurrences of the code S3 in their test history, for example [S3,U1,L2,S3,A1], but as these are not consecutive occurrences, they are *scattered occurrences*.

school group (Ch4)	In order to follow students from primary school to secondary school, schools in the same geographical area were grouped. *School Group A* (SGA), for example, consists of one secondary school and two of its "feeder" primary schools.
school level (Ch4)	*School level* refers to either primary school (up to Grade 6) or secondary school (from Grade 7).
semester 1, 2 (Ch4)	*Semester 1* and *Semester 2* refer to the first and second half of the school year.
sequence of codes (Ch6)	If a student uses the same way of thinking for several consecutive tests, then there will be a *sequence* of equal test codes in their test history. For example, a student with a test history of [L1,S3,S3,A1] has an S3 sequence of length 2. See *scattered occurrences*.
student focussed prevalence (SFP) (Ch5)	The *school level SFP* of a code is calculated by dividing the number of students who were allocated a given code, (on any test that they completed at any time while they were at this particular school level – Primary or Secondary), by the number of students in this particular school level (i.e. the population at risk), expressed as a percent. The *overall SFP* of a code is calculated by dividing the number of students who were allocated this code, (on any test that they completed at any time during the course of the study), by the number of students in the study (expressed as a percent).
superficial teaching (Ch8)	Teaching that focusses on the provision of algorithms rather than improving students' conceptual understanding is referred to as *superficial teaching.*
test (Ch3)	The noun *test* will be used in this thesis to indicate a student's completed script, and DCT will refer to the general test.
test histories (Ch5)	Each student has an (ordered) list of codes indicating the tests that they completed; this is referred to as that student's *test history*. In general a student who has completed three tests has a test history of [T_1,T_2,T_3] but a specific example might be [L,U,A] using the coarse codes or [L1,U1,A2] using the fine codes.

test focussed prevalence (TFP) (Ch5)	The *overall TFP* of a code is calculated by dividing the number of tests allocated this code in this study by the total number of tests completed in this study (expressed as a percent). The *school level TFP* of a code is calculated by dividing the number of tests allocated this code in a given school level (Primary or Secondary) by the total number of tests completed in the same school level (expressed as a percent). The *semester TFP* of a code is calculated by dividing the number of tests allocated this code in a given *semester* (for example semester 2 of Grade 7) by the total number of tests completed in the same *semester* (expressed as a percent).
testing periods (Ch4)	*Testing Periods 1* to *8* refer to the eight occasions over the four years that testing took place. Testing Period 1 (TP1) was in the second half of 1995, and Testing Period 2 was in the first half of 1996, etc.
transitions (Ch5)	Pairs of consecutive tests within student's test histories are examined to provide information about the *transitions* or movements between codes.
truncation (Ch8)	The decimal 4.35 is a *truncation* of 4.3512 as it is created by truncating the longer number.
unstable A1 (Ch8)	The first test in the *regression pair of tests* is referred to as the *unstable A1.*
way of thinking (Ch3)	Distinction is made between the *behaviour* of a student and the *way of thinking* that causes the behaviour.
whole number portion (Ch1)	A number such as 64.530 is said to be composed of two portions: the *whole number portion* (64) and the *decimal portion* (530).

REFERENCES

Af Ekenstam, A. (1977). On Children's Quantitative Understanding of Numbers. *Educational Studies in Mathematics, 8*, 317-332.

Bana, J., Farrell, B., & McIntosh, A. (1997). Student Error Patterns in Fraction and Decimal Concepts. In F. Biddulph & K. Carr (Eds.), *Proceedings of the20th annual conference of the Mathematics Education Research Group of Australasia* (pp. 81-87). Aotearoa, New Zealand: MERGA.

Baturo, A. R., & Cooper, T. (1995). Strategies for Comparing Decimal Numbers With the Same Whole Number Part. In B. Atweh & S. Flavel (Eds.), *Galtha. Proceedings of the 18th annual conference of the Mathematics Education Research Group of Australasia* (pp. 73-79). Darwin: MERGA.

Behr, M. J., & Post, T. R. (1988). Teaching Rational Number and Decimal Concepts. In T. R. Post (Ed.), *Teaching Mathematics in Grades K-8* (pp. 190-231). Boston: Allyn and Bacon.

Behr, M. J., Wachsmuth, I., Post, T., & Lesh, R. (1984). Order and Equivalence of Rational Numbers: A Clinical Teaching Experiment. *Journal for Research in Mathematics Education, 15*(5), 323-341.

Brase, G. (2002). "Bugs" built into the system: How, privileged representations influence mathematical reasoning across the lifespan. *Learning and Individual Differences, 12*(4), 391- 409.

Brekke, G. (1996). A decimal number is a pair of whole numbers. In L. Puig & A. Gutierrez (Eds.), *Proceedings of the 20th conference for the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 137-144). Valencia, Spain: PME.

Brousseau, G. (1997). Problems with Teaching Decimal Numbers, Didactical Problems with Decimals. In N. Balacheff, M. Cooper, R. Sutherland, & V. Warfield (Eds.), *Theory of Didactical Situations in Mathematics : Didactique des Mathematiques, 1970-1990* (pp. 117-232). Dordrecht, The Netherlands: Kluwer Academic.

Brown, J. S., & VanLehn, K. (1982). Towards a Generative Theory of "Bugs". In T. P. Carpenter, J.M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 117-135). Hillsdale, New Jersey: Lawrence Erlbaum.

Brown, M. (1981). Place Value and Decimals. In K. Hart (Ed.), *Children's Understanding of Mathematics: 11-16* (pp. 48-65). London: John Murray.

Brueckner, L. J. (1928). Analysis of Difficulties in Decimals. *Elementary School Journal, 29*, 32-41.

Brueckner, L. J., & Bond, G. (1955). *The diagnosis and treatment of learning difficulties*. New York: Appleton-Century-Crofts.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Lindquist, M. M., & Reys, R. E. (1981). Decimals: Results and implications from national assessment. *Arithmetic Teacher, 21*(8), 34-37.

Clark, H. H., & Clark, E. V. (1977). *Psychology and Language.* New York: Harcourt Brace Jovanovich.

Confrey, J. (1990). A Review of Research on Students Conceptions. In C. Cazden (Ed.), *Review of Research in Education* (Vol. 16, pp. 3-56). Washington, DC: American Educational Research Association.

Foxman, D., Ruddock, G., Joffe, L., Mason, K., Mitchell, P & Sexton, B. (1985). *A Review of Monitoring in Mathematics 1978 to 1982. Assessment of Performance Unit,* (Vol. 1). London: Department of Education and Science.

Fuglestad, A. B. (1998). Computer Support for Diagnostic teaching. A case of decimal numbers. *Nordic Studies in Mathematics Education (Nordisk Matematikk Didaktikk), 6*(3-4), 25-50.

Graeber, A., & Johnson, M. (1991). *Insights into Secondary School Students' Understanding of Mathematics*. College Park, University of Maryland, MD.

Greer, B. (1987). Nonconservation of Multiplication and Division Involving Decimals. *Journal for Research in Mathematics Education, 18*(1), 37-45.

Grossman, A. S. (1983). Decimal Notation: An Important Research Finding. *Arithmetic Teacher, 30*, 32-33.

Helme, S., & Stacey, K. (2000a). Can minimal support for teachers make a difference to students' understanding of decimals? *Mathematics Teacher Education and Development, 2*, 105-120.

Helme, S., & Stacey, K. (2000b). Improving decimal understanding: Can targetted resources make a difference? In J. Bana & A. Chapman (Eds.), *Mathematics Education Beyond 2000, Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia* (pp. 299-306). Fremantle: MERGA.

Hiebert, J. (1986). *Conceptual and Procedural Knowledge: The Case of Mathematics*. Hillsdale, New Jersey: Lawrence Erlbaum.

Hiebert, J. (1987). Research Report: Decimal Fractions. *Arithmetic Teacher, 34*(7), 22-23.

Hiebert, J., & Wearne, D. (1986). Procedures Over Concepts: The Acquisition of Decimal Number Knowledge. In J. Hiebert (Ed.), *Conceptual and Procedural Knowledge: The Case of Mathematics* (pp. 199-223). Hillsdale, New Jersey: Lawrence Erlbaum.

Hiebert, J., Wearne, D., & Taber, S. (1991). Fourth Graders' Gradual Construction of Decimal Fractions During Instruction Using Different Physical Representations. *The Elementary School Journal, 91*(4), 321-341.

Hope, R. A., Longmore, J. M., McManus, S. K., & Wood-Allum., C. A. (1998). *Oxford Handbook of Clinical Medicine. (4th Edition)*: Oxford: Oxford University Press.

Irwin, K. (1996). Making Sense of Decimals. In J. Mulligan & M. Mitchelmore (Eds.), *Children's Number Learning: A Research Monograph of MERGA/AAMT* (pp. 243-257). Adelaide: AAMT.

Irwin, K. (2001). Using Everyday Knowledge of Decimals to Enhance Understanding. *Journal for Research in Mathematics Education, 32*(4), 399-420.

Karch, A., & Karch, F. (2001). The naked decimal point. *American Journal of Nursing, 101*(12), 22.

Kidron, I., & Vinner, S. (1983). Rational Numbers and Decimals at the Senior High Level: Density and Comparison, *Proceedings of the 7th annual conference of the International Group for the Psychology of Mathematics Education* (pp. 301-306). Shoresh, Israel: PME.

Kouba, V. L., Brown, C. A., Carpenter, T. P., Lindquist, M. M., Silver, E. A., & Swafford, J. (1989). Results of the Fourth NAEP Assessment of Mathematics: Number, Operations, and Word Problems. *Arithmetic Teacher, 35*(8), 14-19.

Lesar, T. (2002). Tenfold medication dose prescribing errors. *Annals of Pharmacotherapy, 36*, 1833-1839.

Lightfoot, D. (1999). *Hidden Number Computer Game.* Unpublished Bachelor of Education (Primary) Honours Research Project, University of Melbourne, Melbourne.

MacGregor, M., & Stacey, K. (1997). Students' Understanding of Algebraic Notation 11-15. *Educational Studies in Mathematics, 33*, 1-19.

Markovits, Z., & Sowder, J. (1991). Students' Understanding of the Relationship Between Fractions and Decimals. *Focus on Learning Problems in Mathematics, 13*(1), 3-1.

McIntosh, J., Stacey, K., Tromp, C., & Lightfoot, D. (2000). Designing constructivist computer games for teaching about decimal numbers. In J. Bana & A. Chapman (Eds.), *Mathematics Education Beyond 2000. Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia.* (pp. 409-416). Fremantle: MERGA.

Moloney, K. (1994). *The Evolution of Concepts of Decimals in Primary and Secondary Students.* Unpublished Masters, University of Melbourne, Melbourne.

Moloney, K., & Stacey, K. (1996). Understanding Decimals. *The Australian Mathematics Teacher, 52*(1), 4-8.

Moloney, K., & Stacey, K. (1997). Changes with Age in Students' Conceptions of Decimal Notation. *Mathematics Education Research Journal, 9*(1), 25-38.

Moss, J., & Case, R. (2002). Developing Children's Understanding of the Rational Numbers: A New Model and an Experimental Curriculum. In J. Sowder & B. Schappelle (Eds.), *Lessons Learned From Research* (pp. 143-150): NCTM.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: Mathematics Learning Study Committee, Center for Education, Division of Behavioral Sciences and Education. National Academy Press.

Nesher, P., & Peled, I. (1986). Shifts in Reasoning: The Case of Extending Number Concepts. *Educational Studies in Maths, 17*, 67-79.

Putt, I. J. (1995). Preservice Teachers Ordering of Decimal Numbers: When More is Smaller and Less is Larger! *Focus on Learning Problems in Mathematics, 17*(3), 1-15.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual Bases of Arithmetic Errors: The Case of Decimal Fractions. *Journal for Research in Mathematics Education, 20*(1), 8-27.

Sackur-Grisvard, C., & Leonard, F. (1985). Intermediate Cognitive Organizations in the Process of Learning a Mathematical Concept: The Order of Positive Decimal Numbers. *Cognition and Instruction, 2*(2), 157-174.

Stacey, K., & Flynn, J. (2003). Evaluating an adaptive computer system for teaching about decimals: Two case studies. In V. Aleven, U. Hoppe, J. Kay, R. Mizoguchi, H. Pain, F. Verdejo, & K. Yacef (Eds.), *AI-ED2003 Supplementary Proceedings of the 11th International Conference on Artificial Intelligence in Education.* (pp. 454 – 460). Sydney: University of Sydney. Available at http://www.it.usyd.edu.au/~aied/Supp_procs.html#vol8.

Stacey, K., Helme, S., Archer, S., & Condon, C. (2001). The effect of epistemic fidelity and accessibility on teaching with physical materials: A comparison of two models for teaching decimal numeration. *Educational Studies in Mathematics, 47*, 199-221.

Stacey, K., Helme, S., & Steinle, V. (2001). Confusions between decimals, fractions and negative numbers: A consequence of the mirror as a conceptual metaphor in three different ways. In M. v. d. Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 217-224). Utrecht: PME.

Stacey, K., Helme, S., Steinle, V., Baturo, A., Irwin, K., & Bana, J. (2001). Preservice Teachers' Knowledge of Difficulties in Decimal Numeration. *Journal of Mathematics Teacher Education, 4*(3), 205-225.

Stacey, K., Sonenberg, E., Nicholson, A., Boneh, T. & Steinle, V. (2003). A teaching model exploiting cognitive conflict driven by a Bayesian network. In P. Brusilovsky, A. T. Corbett, and F. De Rosis (Eds.) *Lecture Notes in Artificial Intelligence (Ninth International Conference on User Modeling UM2003) 2702/2003*, 352–362. Springer-Verlag, Heidelberg.

Stacey, K., & Steinle, V. (1998). Refining the Classification of Students' Interpretations of Decimal Notation. *Hiroshima Journal of Mathematics Education, 6*, 49-69.

Stacey, K., & Steinle, V. (1999a). A longitudinal study of children's thinking about decimals: A preliminary analysis. In O. Zaslavsky (Ed.), *Proceedings of the 23rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 233-240). Haifa: PME.

Stacey, K., & Steinle, V. (1999b). Understanding decimals: The path to expertise. In J. M. Truran & K. M. Truran (Eds.), *Making the difference. Proceedings of the 22nd annual conference of the Mathematics Education Research Group of Australasia* (pp. 446-453). Adelaide: MERGA.

Steinle, V., & Stacey, K. (1998a). The incidence of misconceptions of decimal notation amongst students in Grades 5 to 10. In C. Kanes, M. Goos, & E. Warren (Eds.), *Teaching Mathematics in New Times. Proceedings of the 21st annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 548-555). Gold Coast, Australia: MERGA.

Steinle, V., & Stacey, K. (1998b). Students and decimal notation: Do they see what we see? In J. Gough & J. Mousley (Eds.), *35th annual conference of the Mathematics Association of Victoria* (Vol. 1, pp. 415-422). Melbourne: MAV.

Steinle, V., & Stacey, K. (2001). Visible and invisible zeros: Sources of confusion in decimal notation. In J. Bobis, B. Perry & M. Mitchelmore (Eds.), *Numeracy and Beyond. Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 434-441). Sydney: MERGA.

Steinle, V., & Stacey, K. (2002). Further evidence of conceptual difficulties with decimal notation. In B. Barton, K. Irwin, M. Pfannkuch & M. Thomas (Eds.), *Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 633-640). Auckland: MERGA.

Steinle, V., & Stacey, K. (2003a). Exploring the right, probing questions to uncover decimal misconceptions. In L. Bragg, C. Campbell, G. Herbert & J. Mousley (Eds.), *Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 634-641). Geelong: MERGA.

Steinle, V., & Stacey, K. (2003b). Grade-related trends in the prevalence and persistence of decimal misconceptions, *Proceedings of the 27th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 259-266). Hawai'i: PME.

Steinle, V., Stacey, K., & Chambers, D. (2002). *Teaching and Learning about Decimals* [CD-ROM]: Department of Science and Mathematics Education, The University of Melbourne.

Streefland, L. (1991). *Fractions in realistic mathematics education.* Dordrecht: Kluwer Academic.

Swan, M. (1983). *Teaching Decimal Place Value: A Comparative Study of "Conflict" and "Positive Only" Approaches*. Nottingham: Shell Centre for Mathematical Education.

Swan, M. (1990). Becoming Numerate: developing conceptual structures. In S. Willis (Ed.), *Being Numerate: What counts?* Hawthorn, Vic: Australian Council of Educational Research.

Thipkong, S., & Davis, E. J. (1991). Preservice elementary teachers' misconceptions in interpreting and applying decimals. *School Science and Mathematics, 91*(3), 93-99.

TIMSS-R. *IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Retrieved, 29th February, 2004, from the World Wide Web: http://timss.bc.edu/timss1999i/pdf/t99math_items.pdf

Tromp, C. (1999). Number Between: Making a Game of Decimal Numbers. *Australian Primary Mathematics Classroom, 4*(3), 9-11.

Vance, J. (1986). Ordering Decimals and Fractions: A Diagnostic Study. *Focus on Learning Problems in Mathematics, 8*(2), 51-59.

Wearne, D. (1990). Acquiring Meaning for Decimal Fraction Symbols: A One Year Follow-Up. *Educational Studies in Mathematics, 21*, 545-564.

Appendix 1: Decimal Comparison Test Versions

This appendix contains copies of the two versions of the Decimal Comparison Tests (DCT1 and DCT2) used to collect data, as well as 8 tables as follows:

Decimal Quiz Name ___________________

For each pair of decimal numbers, circle whichever is LARGER:

a	4.8	4.63
b	4.08	4.7
c	4.4502	4.45
d	0.5	0.36
e	2.621	2.0687986
f	0.457	0.4
g	0.100	0.25
h	3.72	3.073
i	17.353	17.35
j	0.75	0.8
k	0.038	0.04
l	8.24563	8.245
m	0.37	0.216
n	8.052573	8.514
o	5.62	5.736
p	0.3333333	0.99
q	0.5	0.75
r	1.06	1.053
s	2.06	2.0053
t	0.21	0.021
u	0.3	0.4
v	1.85	1.84
w	6.01	6.1
x	0.25	0.5
y	0.006	0.53

Table 1: Details of the items in DCT1

Item	Decimal pair (larger listed first)		Initial Type	Final Type	Item in DCT2
Q1	4.8	4.63	Type A	Type 1	Q6
Q2	4.7	4.08	Type B	Type 3	Q12
Q3	4.4502	4.45	Type C	Type 4	Q21
Q4	0.5	0.36	Type A	Type 1	Q7
Q5	2.621	2.0687986	Type B	Type 3	Q14
Q6	0.457	0.4	Type C	Type 2*	Q1
Q7	0.25	0.100	Type A	Supplementary	not used
Q8	3.72	3.073	Type B	Type 3	Q13
Q9	17.353	17.35	Type C	Type 4	Q22
Q10	0.8	0.75	Type A	Type 1	Q8
Q11	0.04	0.038	Type B	Supplementary	Q29
Q12	8.24563	8.245	Type C	Type 4	Q23
Q13	0.37	0.216	Type A	Type 1	Q9
Q14	8.514	8.052573	Type B	Type 3	Q15
Q15	5.736	5.62	Type C	Type 2	Q16
Q16	0.99	0.3333333	Supplementary	Supplementary	not used
Q17	0.75	0.5	Supplementary	Type 2	Q17
Q18	1.06	1.053	Supplementary	Supplementary	Q11
Q19	2.06	2.0053	Supplementary	Supplementary	not used
Q20	0.21	0.021	Supplementary	Supplementary	not used
Q21	0.4	0.3	Supplementary	Type 5	Q3
Q22	1.85	1.84	Supplementary	Type 5	Q4
Q23	6.1	6.01	Supplementary	Supplementary	not used
Q24	0.5	0.25	Supplementary	Supplementary	not used
Q25	0.53	0.006	Supplementary	Supplementary	Q30

* Included In DCT2, but not within Type 2

Name:________________________________ Class:____________ Date:____________

For each pair of decimal numbers circle the one which is LARGER.

0.4	0.457	1.06	1.053	4.4502	4.45
0.86	1.3	4.08	4.7	17.353	17.35
0.3	0.4	3.72	3.073	8.245	8.24563
1.85	1.84	2.621	2.0687986	3.2618	3.26
3.71	3.76	8.052573	8.514	3.741	3.746
4.8	4.63	5.62	5.736	0.35	0.42
0.5	0.36	0.5	0.75	2.186	2.954
0.75	0.8	0.426	0.3	0.872	0.813
0.37	0.216	2.516	2.8325	0.038	0.04
3.92	3.4813	7.942	7.63	0.006	0.53

Table 2: Details of the items in DCT2

Item	Decimal pair (larger listed first)		Type	Item in DCT1
Q1	0.457	0.4	Supplementary	Q6
Q2	1.3	0.86	Supplementary	new
Q3	0.4	0.3	Type 5	Q21
Q4	1.85	1.84	Type 5	Q22
Q5	3.76	3.71	Type 5	new
Q6	4.8	4.63	Type 1	Q1
Q7	0.5	0.36	Type 1	Q4
Q8	0.8	0.75	Type 1	Q10
Q9	0.37	0.216	Type 1	Q13
Q10	3.92	3.4813	Type 1	new
Q11	1.06	1.053	Supp	Q18
Q12	4.7	4.08	Type 3	Q2
Q13	3.72	3.073	Type 3	Q8
Q14	2.621	2.0687986	Type 3	Q5
Q15	8.514	8.052573	Type 3	Q14
Q16	5.736	5.62	Type 2	Q15
Q17	0.75	0.5	Type 2	Q17
Q18	0.426	0.3	Type 2	new
Q19	2.8325	2.516	Type 2	new
Q20	7.942	7.63	Type 2	new
Q21	4.4502	4.45	Type 4	Q3
Q22	17.353	17.35	Type 4	Q9
Q23	8.24563	8.245	Type 4	Q12
Q24	3.2618	3.26	Type 4	new
Q25	3.746	3.741	Supplementary	new
Q26	0.42	0.35	Type 6	new
Q27	2.954	2.186	Type 6	new
Q28	0.872	0.813	Type 6	new
Q29	0.04	0.038	Supplementary	Q11
Q30	0.53	0.006	Supplementary	Q25

Table 3: Definitions of item types (larger number listed first)

Item	Samples	Description			
Type	A / B	A is	Ones	Tenths	Comment
1	4.8 / 4.63	shorter	Equal, non-zero	Unequal, non-zero	
2	7.83 / 7.6	longer	Equal, non-zero	Unequal, non-zero	
3	3.6 / 3.07 3.61 / 3.064	shorter	Equal, non-zero	Unequal, A>0, B=0	If ignore 0, then B >A
4	4.5618 / 4.56	longer	Equal, non-zero	Equal, non-zero	Equal, non-zero A has additional digits, but B stops at 2 dp
5	1.4 / 1.3 1.84 / 1.82	Equal length	Equal, non-zero		
6	3.61 / 3.48	Equal length	Equal, non-zero	Unequal, non-zero A<B	Last digit of B>A
7	1.3 / 0.86	shorter	Unequal, A>0, B=0	Unequal, non-zero, A<B	
8	0.6 / 0 0.3 / 0 0.2 / 0	longer B = 0	Both zero	A>0	
9	0.8/0.08 0.21/0.021	shorter	Both zero	Unequal, A>0, B=0	If ignore 0, then A=B
10	3.4/3.40	A=B	Equal	Equal	different lengths

Table 4: Facility on all items in DCT2 by coarse code (n=2331)*

Details			Item (larger listed first)	Sub-type	Overall n=2331	A n=702	L n=649	S n=452	U n=528
Core items	Type1	Q6	4.8 / 4.63	α	64.2	97	7	95	64
		Q7	0.5 / 0.36	β	61.2	97	2	97	55
		Q8	0.8 / 0.75	β	60.5	97	1	96	55
		Q10	3.92 / 3.4813	γ	59.5	94	3	97	51
		Q9	0.37 / 0.216	δ	55.9	92	0	98	40
	Type2	Q17	0.75 / 0.5	β	71.8	96	98	11	60
		Q20	7.942 / 7.63	γ	71.5	98	99	3	61
		Q19	2.8325 / 2.516	γ	70.7	98	98	1	60
		Q16	5.736 / 5.62	γ	70.7	95	99	2	62
		Q18	0.426 / 0.3	δ	70.3	98	100	1	56
Non-Core items	Type3	Q12	4.7 / 4.08	α	73.4	92	34	95	79
		Q15	8.514 / 8.052573	γ	70.9	93	28	96	73
		Q14	2.621 / 2.0687986	γ	69.6	91	25	96	73
		Q13	3.72 / 3.073	γ	69.1	91	25	97	70
	Type4	Q22	17.353 / 17.35	γ	63.5	78	96	6	54
		Q24	3.2618 / 3.26	γ	61.8	75	96	4	52
		Q23	8.24563 / 8.245	γ	60.9	73	94	5	51
		Q21	4.4502 / 4.45	γ	60.3	71	94	5	51
	Type5	Q4	1.85 / 1.84	α	83.3	96	99	45	80
		Q5	3.76 / 3.71	α	82.7	95	99	44	79
		Q3	0.4 / 0.3	β	80.9	94	99	42	75
	Type6	Q26	0.42 / 0.35	β	80.7	98	98	38	74
		Q27	2.954/ 2.186	γ	79.3	98	95	30	76
		Q28	0.872 / 0.813	δ	79.1	98	98	30	73
	Supplementary	Q2	1.3 / 0.86	-	88.1	87	82	93	92
		Q30	0.53 / 0.06	δ	86.7	92	84	89	81
		Q25	3.746 / 3.741	γ	79.3	96	99	31	75
		Q1	0.457 / 0.4	δ	62.0	68	95	8	59
		Q11	1.06 / 1.053	γ	61.6	91	10	95	58
		Q29	0.04 / 0.038	δ	57.9	81	16	90	51

Darker *shaded cells indicate inconsistent responses in core items*
Lighter *shaded cells indicate inconsistent responses in non-core items*
** sample is 2331 tests with maximum score 29/30 completed in 1997*

Table 5: Definition of four sub-types from the interaction of two features

Feature 2: integer part	Feature 1: number of digits after the decimal points	
	Only one or two digits	More than two digits
Non-zero	α	γ
Zero	β	δ

Table 6: Facility on 24 test items (Types 1 to 6) by fine code, Steinle & Stacey (2003a)

Description			Decimal Pair A	Decimal Pair B	Sub-type	Overall n=2331	A1 n=443	A2 n=127	A3 n=132	L1 n=428	L2 n=122	L4 n=99	S1 n=106	S3 n=225	S5 n=121	U1 n=501	U2 n=27
Core items	Type 1	Q6	4.8	4.63	ά	64	97	99	96	4	14	11	97	95	93	67	11
		Q7	0.5	0.36	β	61	97	98	95	2	2	4	95	98	96	57	15
		Q8	0.8	0.75	β	61	98	98	95	1	0	2	95	97	96	57	11
		Q10	3.92	3.4813	γ	60	94	98	88	1	8	8	98	98	96	53	7
		Q9	0.37	0.216	δ	56	91	97	92	0	1	0	95	100	97	42	4
	Type 2	Q17	0.75	0.5	β	72	96	94	95	98	99	97	24	4	13	63	0
		Q20	7.942	7.63	γ	72	100	96	96	100	100	92	2	1	7	64	4
		Q19	2.8325	2.516	γ	71	98	98	97	99	98	95	3	0	1	63	0
		Q16	5.736	5.62	γ	71	95	93	98	99	98	98	4	1	4	65	4
		Q18	0.426	0.3	δ	70	99	97	98	100	100	98	3	0	2	59	0
Non-Core items	Type 3	Q12	4.7	4.08	ά	73	96	98	72	11	95	55	98	96	90	82	19
		Q15	8.514	8.052573	γ	71	96	98	64	3	98	55	99	100	88	76	7
		Q14	2.621	2.0687986	γ	70	96	99	64	2	92	43	97	100	88	76	11
		Q13	3.72	3.073	γ	69	96	99	64	2	92	44	99	99	92	73	7
	Type 4	Q22	17.353	17.35	γ	63	98	9	73	99	100	76	4	1	17	57	7
		Q24	3.2618	3.26	γ	62	96	6	68	100	99	75	2	0	13	55	4
		Q23	8.24563	8.245	γ	61	96	4	64	100	98	67	1	0	17	53	11
		Q21	4.4502	4.45	γ	60	93	8	58	99	94	75	1	0	16	54	0
	Type 5	Q4	1.85	1.84	ά	83	99	100	82	100	99	96	97	4	76	84	0
		Q5	3.76	3.71	ά	83	98	98	83	100	100	95	99	4	70	84	0
		Q3	0.4	0.3	β	81	95	96	87	100	98	95	90	10	60	79	0
	Type 6	Q26	0.42	0.35	β	81	99	97	94	98	99	92	93	8	43	78	7
		Q27	2.954	2.186	γ	79	100	98	94	96	100	86	89	2	31	80	0
		Q28	0.872	0.813	δ	79	100	98	91	100	100	87	92	2	29	77	0

Table 7: Facility on 6 supplementary items by fine code, (unpublished)

Item (larger listed first)		Sub-type	Overall n=2331	A1 (n=443)	A2 (n=127)	A3 (n=132)	L1 (n=428)	L2 (n=122)	L4 (n=99)	S1 (n=106)	S3 (n=225)	S5 (n=121)	U1 (n=501)	U2 (n=27)
Q2	1.3 / 0.86	-	88	85	92	89	78	95	84	99	90	93	94	67
Q30	0.53 / 0.006	δ	87	93	95	86	78	98	89	94	91	82	85	7
Q25	3.746 / 3.741	γ	79	98	94	90	99	99	96	86	4	33	79	0
Q1	0.457 / 0.4	δ	62	80	35	61	96	92	93	16	2	12	62	0
Q11	1.06 / 1.053	γ	62	91	94	87	3	24	22	93	99	91	60	4
Q29	0.04 / 0.038	δ	58	79	93	78	7	41	29	85	96	84	53	4

Table 8: Details of the items in 15-item test (Moloney & Stacey, 1997)

Item	Decimal pair (larger listed first)		Type
Q1	4.8	4.63	Type A
Q2	0.4	0.36	Type A
Q3	0.35	0.100	Type A
Q4	0.8	0.75	Type A
Q5	0.37	0.216	Type A
Q6	4.7	4.08	Type B
Q7	2.621	2.0687986	Type B
Q8	3.72	3.073	Type B
Q9	0.2	0.038	Type B
Q10	8.514	8.0525738	Type B
Q11	4.4502	4.45	Type C
Q12	0.457	0.4	Type C
Q13	17.353	17.35	Type C
Q14	8.24563	8.245	Type C
Q15	5.736	5.62	Type C

Appendix 1

Appendix 2: Pilot study on arrangement of test items

Two versions of a 25-item Decimal Comparison Test were trialled with a sample of 163 Year 7 students. This pilot study was conducted to determine whether the order that the items were presented on the test affected student's responses. These two versions appear on the next two pages; in the first version the items are *ordered* according to the item types, and in the second, the item types are *mixed* throughout the test. Note that the first question on both test versions is a "warm-up" item, and did not contribute to the allocation of codes to tests. The five tables included in this appendix are listed below:

Table 1: Item types and question number for the two test versions
Table 2: Number of students allocated to fine codes by test version
Table 3: Analysis of variance for frequencies of fine codes
Table 4: Number of deviations from predictions on all test items by test version
Table 5: Two-way analysis of variance for deviations

Table 1: Item types and question number for the two test versions

Item (larger listed first)		Item Type	Test version	
			Ordered	Mixed
4.8	4.63	1	Q2	Q2
0.5	0.36	1	Q3	Q8
0.8	0.75	1	Q4	Q14
0.37	0.216	1	Q5	Q20
3.92	3.4813	1	Q6	Q24
5.736	5.62	2	Q7	Q3
0.75	0.5	2	Q8	Q9
0.426	0.3	2	Q9	Q15
2.8325	2.516	2	Q10	Q21
7.942	7.63	2	Q11	Q25
4.7	4.08	3	Q12	Q4
3.72	3.073	3	Q13	Q10
2.621	2.0687986	3	Q14	Q16
8.514	8.052573	3	Q15	Q22
4.4502	4.45	4	Q16	Q5
17.353	17.35	4	Q17	Q11
8.24563	8.245	4	Q18	Q17
3.2618	3.26	4	Q19	Q23
0.4	0.3	5	Q20	Q6
1.85	1.84	5	Q21	Q12
3.746	3.741	5	Q22	Q18
0.42	0.35	6	Q23	Q7
2.954	2.186	6	Q24	Q13
0.872	0.813	6	Q25	Q19

It is hypothesised that in the *ordered* test version, a student might notice some similarity between the current item and the items immediately preceding, and this may affect their choice on the current item. In the *mixed* test version, however, it is more unlikely that students would notice any similarities between items. Hence, it is predicted that the *mixed* test version would generate more inconsistent choices with one of the following results;

- enough inconsistent choices to be allocated U1, or
- a few inconsistent choices which does not affect their classification.

Two analyses of these inconsistencies were conducted. The first analysis is to determine whether the test version (ordered or mixed) affects the frequency of the various codes allocated to students' tests, while the second considers the *deviations* from the predictions made on each item.

Decimal Quiz Name:_______________________

Class: _________________

For each pair of decimal numbers, circle whichever is LARGER:

1)	0.4	0.457
2)	4.8	4.63
3)	0.5	0.36
4)	0.75	0.8
5)	0.37	0.216
6)	3.92	3.4813
7)	5.62	5.736
8)	0.5	0.75
9)	0.426	0.3
10)	2.516	2.8325
11)	7.63	7.942
12)	4.08	4.7
13)	3.72	3.073
14)	2.621	2.0687986
15)	8.052573	8.514
16)	4.4502	4.45
17)	17.353	17.35
18)	8.245	8.24563
19)	3.2618	3.26
20)	0.3	0.4
21)	1.85	1.84
22)	3.741	3.746
23)	0.35	0.42
24)	2.186	2.954
25)	0.813	0.872

Decimal Quiz Name:______________________
Class: _________________

For each pair of decimal numbers, circle whichever is LARGER:

1)	0.4	0.457
2)	4.8	4.63
3)	5.62	5.736
4)	4.08	4.7
5)	4.4502	4.45
6)	0.3	0.4
7)	0.35	0.42
8)	0.36	0.5
9)	0.5	0.75
10)	3.72	3.073
11)	17.353	17.35
12)	1.85	1.84
13)	2.186	2.954
14)	0.75	0.8
15)	0.426	0.3
16)	2.621	2.0687986
17)	8.245	8.24563
18)	3.741	3.746
19)	0.813	0.872
20)	0.37	0.216
21)	2.516	2.8325
22)	8.052573	8.514
23)	3.2618	3.26
24)	3.92	3.4813
25)	7.63	7.942

Analysis 1: Frequency of codes by test version

Table 2 provides the number of students allocated to each of the 12 fine codes according to test version. (Note that the allocation of fine codes to students' test papers is explained fully in Chapter 3 and that the code S2 was not used in 1997). The distribution of the 82 ordered tests appears very similar to the distribution of the 81 mixed tests.

Table 2: Number of students allocated to fine codes by test version

Fine Code	Test Version	
	Ordered	Mixed
A1	23	22
A2	4	3
A3	3	6
L1	11	11
L2	3	1
L3	1	1
L4	1	1
S1	3	0
S3	10	11
S4	4	1
U1	19	22
U2	0	2
Total	82	81

An analysis of variance was conducted to determine if the frequencies of the fine codes were affected by test version. Conclusion: test version is not statistically significant in the model (p=0.891).

Table 3: Analysis of Variance for frequencies of fine codes

Source	DF	SS	MS	F	P
fine code	11	1304.46	118.59	55.61	0.000
test version	1	0.04	0.04	0.02	0.891
Error	11	23.46	2.13		
Total	23	1327.96			

The data was reanalysed using a Log Linear Model, with the fine codes being grouped. Rather than grouping into the coarse codes (A, L, S & U), it was decided that it was more appropriate to combine all the various "unclassified codes". Hence, the four codes used in this analysis were:

A'=A1+A2, L'=L1+L2+L3, S'=S1+S3, U'=A3+L4+ S4+U1+U2

While there were more test papers in the U’ group for the mixed test version (32 cf 27), the new analysis confirmed that test version was not significant (p=0.938).

Analysis 2: Deviations from predictions by test version

If a student had only some inconsistent responses, they may still be allocated to one of the misconceptions; these inconsistencies will be observed as *deviations* from the predictions for that particular misconception. Note that this analysis of deviations only makes sense for the seven fine codes that have predictions for all 24 items (i.e. A1, A2, L1, L2 L3, S1 and S3).

Table 4: Number of deviations from predictions on all test items by test version

Fine Code	Deviations	
	Ordered	Mixed
A1	7	8
A2	4	6
L1	9	14
L2	1	3
L3	3	1
S1	3	0
S3	8	7
Total	35	39

An analysis of variance was conducted to determine if the number of deviations from predictions was affected by test version. Conclusion: that test version is not statistically significant in the model (p=0.604).

Table 5: Two-way analysis of variance for deviations

Source	DF	SS	MS	F	P
code	6	168.86	28.14	7.39	0.014
test version	1	1.14	1.14	0.30	0.604
Error	6	22.86	3.81		
Total	13	192.86			

An alternative analysis using the Log Linear Model was conducted by grouping the fine codes into three coarse codes (A’, L’ and S’) as previously defined. This new analysis confirmed that test version was not significant (p=0.642).

Appendix 3: Meta-data from the ARC study

Table 1: Number of students with first and last tests in the primary sSample

Semester of first test	Semester of last test						Total
	Grade 4		Grade 5		Grade 6		
Gr4-Sem1	18	0	1	0	3	17	39
Gr4-Sem2		36	5	3	22	231	297
Gr5-Sem1			45	11	16	197	269
Gr5-Sem2				30	1	179	210
Gr6-Sem1					108	41	149
Gr6-Sem2						115	115
Total	18	36	51	44	150	780	1079

Table 2: Number of students with first and last tests in the secondary sample

Semester of first test	Semester of last test								Total
	Grade 7		Grade 8		Grade 9		Grade 10		
Gr7-Sem1	24	13	10	57	275	37			416
Gr7-Sem2		94	29	114	216	73	153	52	731
Gr8-Sem1			10	0	1	9	9	40	69
Gr8-Sem2				81	19	72	84	60	316
Gr9-Sem1					5	0	0	17	22
Gr9-Sem2						106	10	172	288
Gr10-Sem1							38	0	38
Gr10-Sem2								245	245
Total	24	107	49	252	516	297	294	586	2125

Table 3: Numbers of tests in each semester by school group

	Semester	School Group						Total
		SGA	SGB	SGC	SGD	SGE	SGF	
Primary	Gr4-Sem1	39						39
	Gr4-Sem2	0	29	178	90			297
	Gr5-Sem1	125	0	93	157	39		414
	Gr5-Sem2	33	66	262	104	82		547
	Gr6-Sem1	143	27	188	227	84		669
	Gr6-Sem2	89	113	266	190	122		780
	subtotal	*429*	*235*	*987*	*768*	*327*	*0*	*2746*
Secondary	Gr7-Sem1	127	23	346	285	130		911
	Gr7-Sem2	202	333	269	193	184	205	1386
	Gr8-Sem1	130	33	263	193	167	131	917
	Gr8-Sem2	215	294	196	104	55	325	1189
	Gr9-Sem1	102	95	196	111	122	235	861
	Gr9-Sem2	188	241			58	299	786
	Gr10-Sem1	133	92			50	205	480
	Gr10-Sem2	145	157				284	586
	subtotal	*1242*	*1268*	*1270*	*886*	*766*	*1684*	*7116*
	Total	1671	1503	2257	1654	1093	1684	9862

Table 4: Average number of previous tests in each semester by school group

Semester	School Group						Overall
	SGA	SGB	SGC	SGD	SGE	SGF	
Gr4-Sem1	0.0						0.0
Gr4-Sem2		0.0	0.0	0.0			0.0
Gr5-Sem1	0.2		0.9	0.3	0.0		0.4
Gr5-Sem2	1.2	0.3	0.8	0.9	0.5		0.8
Gr6-Sem1	0.9	1.8	2.1	1.2	1.3		1.4
Gr6-Sem2	2.3	1.0	2.3	2.3	1.5		2.0
Gr7-Sem1	0.8	3.7	1.4	1.9	1.9		1.6
Gr7-Sem2	0.6	0.3	1.7	2.2	1.2	0.0	1.0
Gr8-Sem1	0.6	2.7	2.6	3.0	2.2	1.0	2.1
Gr8-Sem2	1.0	0.8	3.0	2.9	3.1	0.9	1.5
Gr9-Sem1	1.2	2.2	3.7	3.8	2.8	2.1	2.7
Gr9-Sem2	1.2	0.7			2.8	1.0	1.1
Gr10-Sem1	1.2	2.2			3.6	2.1	2.0
Gr10-Sem2	1.6	0.7				1.1	1.1

Empty cell indicates no testing and is not the same as 0.0

Table 5: School Group A- Tests in each semester by cohort

Semester	Cohort (number of students)							Total (n=709)
	1992 (n=72)	1991 (n=82)	1990 (n=147)	1989 (n=146)	1988 (n=160)	1987 (n=64)	1986 (n=38)	
Gr4-Sem1	39							39
Gr4-Sem2	0							0
Gr5-Sem1	53	72						125
Gr5-Sem2	33	0						33
Gr6-Sem1	40	49	54					143
Gr6-Sem2	44	45	0					89
Gr7-Sem1		45	0	82				127
Gr7-Sem2		31	123	0	48			202
Gr8-Sem1			54	0	76			130
Gr8-Sem2			90	125	0			215
Gr9-Sem1				80	0	22		102
Gr9-Sem2				65	123	0		188
Gr10-Sem1					95	0	38	133
Gr10-Sem2					86	59	0	145
Total	209	242	321	352	428	81	38	1671

Table 6: School Group B- Tests in each semester by cohort

Semester	Cohort (number of students)							Total (n=679)
	1992 (n=33)	1991 (n=45)	1990 (n=147)	1989 (n=140)	1988 (n=138)	1987 (n=130)	1986 (n=46)	
Gr4-Sem1								
Gr4-Sem2	29							29
Gr5-Sem1	0							0
Gr5-Sem2	26	40						66
Gr6-Sem1	27	0						27
Gr6-Sem2	27	41	45					113
Gr7-Sem1	23	0	0					23
Gr7-Sem2		34	135	90	74			333
Gr8-Sem1		33	0	0	0			33
Gr8-Sem2			112	119	39	24		294
Gr9-Sem1			95	0	0	0		95
Gr9-Sem2				76	88	47	30	241
Gr10-Sem1				92	0	0	0	92
Gr10-Sem2					53	85	19	157
Total	132	148	387	377	254	156	49	1503

Table 7: School Group C- Tests in each semester by cohort

Semester	Cohort (number of students)							Total (n=549)
	1992 (n=107)	1991 (n=117)	1990 (n=325)	1989 (n=0)	1988 (n=0)	1987 (n=0)	1986 (n=0)	
Gr4-Sem1								
Gr4-Sem2	95	83						178
Gr5-Sem1	93	0						93
Gr5-Sem2	87	92	83					262
Gr6-Sem1	91	97	0					188
Gr6-Sem2	91	90	85					266
Gr7-Sem1	55	41	250					346
Gr7-Sem2		37	232					269
Gr8-Sem1		46	217					263
Gr8-Sem2			196					196
Gr9-Sem1			196					196
Gr9-Sem2								
Gr10-Sem1								
Gr10-Sem2								
Total	512	486	1259					2257

Table 8: School Group D- Tests in each semester by cohort

Semester	Cohort (number of students)							Total (n=397)
	1992 (n=113)	1991 (n=129)	1990 (n=155)	1989 (n=0)	1988 (n=0)	1987 (n=0)	1986 (n=0)	
Gr4-Sem1								
Gr4-Sem2		90						90
Gr5-Sem1	102	55						157
Gr5-Sem2	104	0						104
Gr6-Sem1	84	95	48					227
Gr6-Sem2	93	97	0					190
Gr7-Sem1	76	79	130					285
Gr7-Sem2		75	118					193
Gr8-Sem1		69	124					193
Gr8-Sem2			104					104
Gr9-Sem1			111					111
Gr9-Sem2								
Gr10-Sem1								
Gr10-Sem2								
Total	459	560	635					1654

Table 9: School Group E- Tests in each semester by cohort

Semester	Cohort (number of students)							Total (n=258)
	1992 (n=43)	1991 (n=53)	1990 (n=76)	1989 (n=86)	1988 (n=0)	1987 (n=0)	1986 (n=0)	
Gr4-Sem1								
Gr4-Sem2								
Gr5-Sem1	39							39
Gr5-Sem2	43	39						82
Gr6-Sem1	39	45						84
Gr6-Sem2	40	45	37					122
Gr7-Sem1	29	41	60					130
Gr7-Sem2		41	60	83				184
Gr8-Sem1		39	55	73				167
Gr8-Sem2			55	0				55
Gr9-Sem1			54	68				122
Gr9-Sem2				58				58
Gr10-Sem1				50				50
Gr10-Sem2								
Total	190	250	321	332				1093

Table 10: School Group F- Tests in each semester by cohort

Semester	Cohort (number of students)							Total (n=612)
	1992 (n=0)	1991 (n=0)	1990 (n=146)	1989 (n=149)	1988 (n=146)	1987 (n=148)	1986 (n=23)	
Gr4-Sem1								
Gr4-Sem2								
Gr5-Sem1								
Gr5-Sem2								
Gr6-Sem1								
Gr6-Sem2								
Gr7-Sem1								
Gr7-Sem2			146	59				205
Gr8-Sem1			131	0				131
Gr8-Sem2			125	139	61			325
Gr9-Sem1			124	111	0			235
Gr9-Sem2				112	139	48		299
Gr10-Sem1				83	122	0		205
Gr10-Sem2					122	139	23	284
Total			526	504	444	187	23	1684

Appendix 3

Appendix 4: Test-focussed prevalence of codes

The tables in this appendix contain the test-focussed prevalence (TFP) of each of the codes by semester and school group. Whilst these tables accurately report this prevalence, not attempt is made to remove the project effect. Table 1 contains the sample size for the cells in Tables 2 to 18, with the seven samples of less than 50 being italicised for easy reference. Tables 2 to 18 are arranged by coarse code (A, L, S, U) and by fine code. The last row of the tables for the fine codes indicates the proportion of the coarse code that this fine code contributes each semester. Tables 17 and 18 contain the TFP of combinations of two fine codes; A2 and A3 (i.e. A but not A1) and L2 and L4 (i.e. L but not L1). Table 19 contains the TFP of the codes when only first tests are considered; the adjustments to A1 and A due to SGE not being included in Grades 8, 9 and 10 are then listed.

Table 1: Number of tests by semester and by school group

School Group	Primary School Level (4-6)						Secondary School Level (7-10)								Primary Total	Secondary Total
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2	Sem 1	Sem 2		
SGA	*39*		125	*33*	143	89	127	202	130	215	102	188	133	145	429	1242
SGB		*29*		66	*27*	113	*23*	333	*33*	294	95	241	92	157	235	1268
SGC		178	93	262	188	266	346	269	263	196	196				987	1270
SGD		90	157	104	227	190	285	193	193	104	111				768	886
SGE			*39*	82	84	122	130	184	167	55	122	58	50		327	766
SGF								205	131	325	235	299	205	284	0	1684
Total	39	297	414	547	669	780	911	1386	917	1189	861	786	480	586	2746	7116

Cells containing sample numbers of smaller than 50 have been italicised. This formatting has been retained in Tables 2–18 to draw attention to any small samples.

Table 2: Test-focussed prevalence (TFP) of A by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		7	*12*	13	24	37	59	43	48	56	64	66	72	12	56
SGB		*7*		21	*63*	41	*78*	62	*67*	65	84	65	82	78	34	69
SGC		6	6	24	30	59	62	67	74	72	70				30	68
SGD		9	31	30	56	71	73	70	71	74	74				46	72
SGE			*28*	62	81	74	82	84	85	91	89	90	98		67	87
SGF								35	79	74	79	75	80	73		71
Overall	**0**	**7**	**18**	**30**	**43**	**58**	**65**	**63**	**71**	**68**	**76**	**70**	**79**	**74**	**36**	**69**

Table 3: Test-focussed prevalence (TFP) of A1 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		6	*9*	9	13	29	50	35	44	51	54	55	66	8	48
SGB		*3*		17	*59*	35	*65*	54	*64*	56	79	58	72	71	29	61
SGC		3	2	18	24	52	53	60	67	67	64				24	61
SGD		6	25	23	49	66	66	61	67	64	67				40	65
SGE			*26*	62	79	67	79	77	78	87	85	86	90		64	81
SGF								29	73	61	66	61	70	62		60
Overall	**0**	**4**	**14**	**25**	**38**	**51**	**58**	**55**	**65**	**59**	**68**	**60**	**68**	**65**	**31**	**61**
Proportion of A		60	77	83	88	88	89	88	91	88	90	86	87	88	86	88

Table 4: Test-focussed prevalence (TFP) of A2 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		0	*3*	3	6	6	6	5	2	2	5	8	2	2	5
SGB		*0*		3	*0*	5	*0*	5	*0*	5	2	4	3	4	3	4
SGC		1	2	3	3	3	4	3	5	4	5				2	4
SGD		2	3	5	4	3	5	6	2	8	3				3	4
SGE			*0*	0	2	4	2	5	5	2	3	2	2		2	4
SGF								4	4	8	6	10	7	6		7
Overall	**0**	**1**	**2**	**3**	**3**	**4**	**4**	**5**	**4**	**5**	**4**	**6**	**6**	**5**	**3**	**5**
Proportion of A		15	9	10	7	6	6	8	6	8	5	9	8	6	8	7

Table 5: Test-focussed prevalence (TFP) of A3 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		2	*0*	1	4	2	2	3	2	3	5	4	4	2	3
SGB		*3*		2	*4*	1	*13*	3	*3*	4	3	3	7	3	2	4
SGC		2	2	3	3	5	5	4	2	1	2				3	3
SGD		1	3	2	3	2	2	3	3	2	5				2	3
SGE			*3*	0	0	2	1	2	2	2	1	2	6		1	2
SGF								1	2	5	6	4	3	5		4
Overall	**0**	**2**	**2**	**2**	**2**	**3**	**3**	**3**	**2**	**3**	**3**	**4**	**4**	**4**	**2**	**3**
Proportion of A		25	13	7	5	6	5	4	3	5	5	5	5	6	7	5

Table 6: Test-focussed prevalence (TFP) of L by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*85*		83	*55*	59	37	37	20	21	18	10	7	11	8	64	16
SGB		*90*		36	*7*	18	*4*	11	*6*	9	6	7	3	3	31	7
SGC		70	59	51	38	18	18	12	6	6	4				44	10
SGD		73	34	26	17	12	9	4	5	5	4				27	6
SGE			*21*	7	7	15	5	2	1	0	0	0	0		12	2
SGF								21	3	5	1	3	0	2		5
Overall	**85**	**73**	**53**	**38**	**30**	**18**	**16**	**12**	**6**	**8**	**4**	**5**	**4**	**4**	**37**	**8**

Table 7: Test-focussed prevalence (TFP) of L1 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*82*		77	*55*	46	24	28	13	8	12	5	4	5	6	54	10
SGB		*83*		24	*4*	13	*4*	6	*0*	4	4	3	3	1	24	4
SGC		57	43	34	24	9	12	7	3	2	3				30	6
SGD		66	20	15	8	8	6	3	2	1	1				18	3
SGE			*18*	5	4	4	2	1	1	0	0	0	0		6	1
SGF								11	2	2	1	0	0	1		2
Overall	**82**	**62**	**42**	**26**	**20**	**10**	**10**	**7**	**3**	**4**	**2**	**2**	**2**	**2**	**27**	**4**
Proportion of L	97	85	79	69	67	57	66	58	41	51	55	40	53	50	73	55

Table 8: Test-focussed prevalence (TFP) of L2 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*3*		2	*0*	8	6	4	4	9	4	3	2	2	1	5	4
SGB		*0*		8	*0*	2	*0*	3	*3*	3	0	3	0	1	3	2
SGC		6	10	8	7	4	5	3	2	4	1				6	3
SGD		4	6	5	5	3	2	1	2	2	3				5	2
SGE			*0*	1	2	6	2	1	1	0	0	0	0		3	1
SGF								5	2	2	0	2	0	1		2
Overall	**3**	**5**	**5**	**6**	**6**	**4**	**3**	**3**	**3**	**3**	**1**	**2**	**1**	**1**	**5**	**2**
Proportion of L	3	6	10	15	19	22	22	24	41	32	29	43	18	27	14	28

Table 9: Test-focussed prevalence (TFP) of L4 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		4	*0*	6	8	5	3	4	2	2	1	4	1	5	3
SGB		*7*		5	*4*	3	*0*	2	*3*	2	2	1	0	1	4	1
SGC		8	6	9	6	5	2	2	2	1	1				7	2
SGD		3	8	6	3	1	1	1	1	2	0				4	1
SGE			*3*	1	1	5	1	1	0	0	0	0	0		3	0
SGF								5	0	1	0	1	0	1		1
Overall	**0**	**6**	**6**	**6**	**4**	**4**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**5**	**1**
Proportion of L	0	9	11	16	14	21	12	19	19	17	16	18	29	23	13	17

Table 10: Test-focussed prevalence (TFP) of S by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*5*		5	*15*	9	11	13	9	22	15	14	14	8	7	8	1
SGB		*0*		30	*26*	32	*13*	18	*12*	14	5	14	8	7	27	13
SGC		12	20	12	14	11	9	6	10	13	8				13	9
SGD		7	24	16	15	7	8	14	14	13	11				14	12
SGE			*10*	11	2	6	8	11	8	4	5	5	0		7	7
SGF								30	7	11	8	8	8	9		11
Overall	**5**	**9**	**16**	**15**	**13**	**12**	**9**	**15**	**12**	**13**	**8**	**11**	**7**	**8**	**13**	**11**

Table 11: Test-focussed prevalence (TFP) of S1 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		1	*6*	3	2	4	5	4	0	2	2	2	0	2	2
SGB		*0*		8	*4*	10	*9*	7	*0*	3	3	4	1	3	7	4
SGC		3	4	5	9	3	3	2	1	3	2				5	2
SGD		1	9	6	6	4	3	5	2	3	3				5	3
SGE			*3*	4	1	2	2	4	1	0	1	2	0		2	2
SGF								6	2	2	3	3	2	1		3
Overall	**0**	**2**	**5**	**5**	**5**	**4**	**3**	**5**	**2**	**2**	**2**	**3**	**2**	**1**	**4**	**3**
Proportion of S	0	26	30	34	42	31	33	35	16	16	28	25	24	13	34	25

Table 12: Test-focussed prevalence (TFP) of S3 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		3	*9*	2	4	3	1	13	10	7	6	2	7	3	6
SGB		*0*		15	*19*	14	*4*	6	*12*	6	1	5	2	2	13	5
SGC		5	11	5	4	5	4	1	6	8	4				5	4
SGD		2	10	10	6	3	4	7	8	9	7				6	6
SGE			*3*	4	1	2	4	4	4	4	4	3	0		2	4
SGF								17	5	6	2	3	3	4		5
Overall	**0**	**4**	**7**	**7**	**4**	**5**	**4**	**6**	**7**	**7**	**4**	**5**	**3**	**4**	**6**	**5**
Proportion of S	0	41	47	47	35	44	42	41	60	56	48	42	36	50	42	48

Table 13: Test-focussed prevalence (TFP) of S5 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*5*		1	*0*	4	4	6	2	5	4	5	6	3	0	3	4
SGB		*0*		8	*4*	8	*0*	5	*0*	5	1	5	4	3	6	4
SGC		3	5	3	2	3	2	2	2	3	2				3	2
SGD		3	4	1	4	1	2	2	4	1	1				3	2
SGE			*5*	4	0	2	2	3	3	0	0	0	0		2	2
SGF								7	0	3	3	2	2	5		3
Overall	**5**	**3**	**4**	**3**	**3**	**3**	**2**	**4**	**3**	**4**	**2**	**4**	**3**	**3**	**3**	**3**
Proportion of S	100	33	23	19	24	25	25	24	24	28	24	34	39	37	24	27

Table 14: Test-focussed prevalence (TFP) of U by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*10*		5	*18*	18	28	13	11	14	19	21	14	16	14	16	15
SGB		*3*		12	*4*	10	*4*	10	*15*	12	4	14	8	12	9	11
SGC		12	14	13	18	12	11	14	11	9	17				14	12
SGD		11	11	28	12	9	9	12	10	9	12				13	10
SGE			*41*	20	10	6	5	2	6	5	6	5	2		14	5
SGF								13	11	10	12	14	12	16		13
Overall	**10**	**11**	**13**	**17**	**14**	**12**	**10**	**11**	**10**	**12**	**13**	**13**	**11**	**15**	**14**	**12**

Table 15: Test-focussed prevalence (TFP) of U1 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*10*		4	*18*	18	28	13	11	12	16	20	13	14	10	15	14
SGB		*3*		12	*4*	10	*4*	8	*9*	11	4	12	7	10	9	9
SGC		12	14	13	17	12	11	14	10	8	15				13	11
SGD		11	11	28	12	9	9	12	10	5	11				13	10
SGE			*41*	20	10	6	5	2	4	4	6	3	2		14	4
SGF								13	8	9	11	12	8	15		11
Overall	**10**	**11**	**13**	**17**	**14**	**12**	**10**	**10**	**9**	**10**	**11**	**12**	**9**	**13**	**13**	**10**
Proportion of U	100	100	98	100	99	99	100	95	85	86	90	88	79	88	99	90

Table 16: Test-focussed prevalence (TFP) of U2 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		1	*0*	0	0	0	0	2	3	1	1	2	3	0	1
SGB		*0*		0	*0*	0	*0*	1	*6*	1	0	2	1	2	0	1
SGC		0	0	0	1	0	0	1	1	2	3				0	1
SGD		0	0	0	0	0	0	0	1	4	1				0	1
SGE			*0*	0	0	0	0	1	2	2	0	2	0		0	1
SGF								0	2	1	2	2	4	1		1
Overall	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**2**	**1**	**2**	**2**	**2**	**0**	**1**
Proportion of U	0	0	2	0	1	1	0	5	15	14	10	12	21	12	1	10

Table 17: Test-focussed prevalence (TFP) of A2+A3 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*0*		2	*3*	4	10	8	9	8	4	5	10	11	6	4	8
SGB		*3*		5	*4*	6	*13*	8	*3*	9	5	7	10	7	5	8
SGC		2	4	6	6	8	8	7	7	5	7				6	7
SGD		3	6	7	7	5	7	9	4	10	7				6	7
SGE			*3*	0	2	7	3	7	7	4	4	3	8		3	5
SGF								6	6	13	12	14	10	11		11
Overall	**0**	**3**	**4**	**5**	**5**	**7**	**7**	**8**	**6**	**8**	**8**	**10**	**10**	**9**	**5**	**8**
Proportion of A	0	40	23	17	12	12	11	12	9	12	10	14	13	12	14	12

Table 18: Test-focussed prevalence (TFP) of L2+L4 by semester and by school group

School Group	Semester														Prim	Sec
	Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10			
SGA	*3*		6	*0*	13	13	9	7	13	6	5	3	6	2	9	6
SGB		*7*		12	*4*	4	*0*	5	*6*	5	2	4	0	2	7	4
SGC		13	16	16	13	9	7	5	3	4	2				13	4
SGD		8	14	11	8	4	3	2	3	4	3				9	3
SGE			*3*	2	4	11	3	2	1	0	0	0	0		6	1
SGF								10	2	2	0	3	0	2		3
Overall	**3**	**11**	**11**	**12**	**10**	**8**	**5**	**5**	**4**	**4**	**2**	**3**	**2**	**2**	**10**	**4**
Proportion of L	3	15	21	31	33	43	34	42	59	49	45	60	47	50	27	45

Table 19: Test-focussed prevalence (TFP) of all codes by grade for first tests only*

Codes	Grade 4 (n=336)	Grade 5 (n=479)	Grade 6 (n=264)	Grade 7 (n=1147)	Grade 8 (n=385)	Grade 9 (n=310)	Grade 10 (n=283)
A1	3.6	24.0	31.1	47.6	43.1	54.8	60.4
A2	0.9	1.9	7.2	5.7	6.5	7.1	6.0
A3	1.5	1.7	2.3	2.5	3.1	3.2	3.5
Total A	**6.0**	**27.6**	**40.5**	**55.8**	**52.7**	**65.2**	**70.0**
L1	64.3	35.3	19.3	10.3	8.1	2.9	2.8
L2	4.5	4.8	4.9	3.7	4.2	1.6	1.4
L4	5.7	5.6	8.3	3.1	3.1	1.3	2.1
Total L	**74.4**	**45.7**	**32.6**	**17.1**	**15.3**	**5.8**	**6.4**
S1	2.1	4.2	6.1	5.2	2.6	3.9	1.8
S3	3.3	6.1	5.3	7.1	9.4	3.9	3.9
S5	3.3	4.2	7.2	4.6	6.0	5.2	4.2
Total S	**8.6**	**14.4**	**18.6**	**17.0**	**17.9**	**12.9**	**9.9**
U1	11.0	12.1	8.3	9.8	11.9	14.5	13.1
U2	0.0	0.2	0.0	0.3	2.1	1.6	0.7
Total U	**11.0**	**12.3**	**8.3**	**10.1**	**14.0**	**16.1**	**13.8**

*Adjustments made to the shaded figures in the reporting in Chapter 9 to compensate for the absence of SGE in this data: Prevalence of A1 in Grades 8, 9, 10: 60%, 65% and 70%, respectively, and prevalence of A is then 70%, 75% and 80% respectively.

Appendix 4

Appendix 5: Student-focussed prevalence of codes for various samples

This appendix contains an analysis of the effect on the SFP of the codes by removing the less-tested students from the samples. Each of the 3204 students was allocated to a sample depending on the grade of their first test. Table 1 indicates the total number of tests completed by each of the 1079 students in the primary sample, and the 2125 students in the secondary sample. For example, 65 students completed exactly 5 tests in primary school; some of these students were followed to secondary school but any additional tests completed there do not contribute to the determination of the primary level SFP.

Table 1: Distribution of the total number of tests by a student for the primary and secondary Samples

Total number of tests for a student	Primary sample	Secondary sample
1	352	603
2	185	388
3	209	452
4	268	443
5	65	239
Total	1079	2125

The SFP for the coarse and fine codes will be determined on various samples that have been created by imposing restrictions of each of the samples, to remove students who have completed fewer tests. These sample sizes are provided in Table 2; for example the last row indicates that there were over 300 students who completed over 4 tests in primary school and nearly 700 students who completed over 4 tests in secondary school.

Table 2: Numbers of students in various restricted samples

Total number of tests for a student	Primary sample	Secondary sample
1 to 5 tests	1079	2125
2 to 5 tests	727	1522
3 to 5 tests	542	1134
4 to 5 tests	333	682

The general increasing trends in Figures 1 and 2 indicate that the initial calculations for SFP are underestimates due to students who are less-tested diluting the sample. Due to this, the SFP for the most restricted samples (4 or 5 tests) will be presented in Chapter 9.

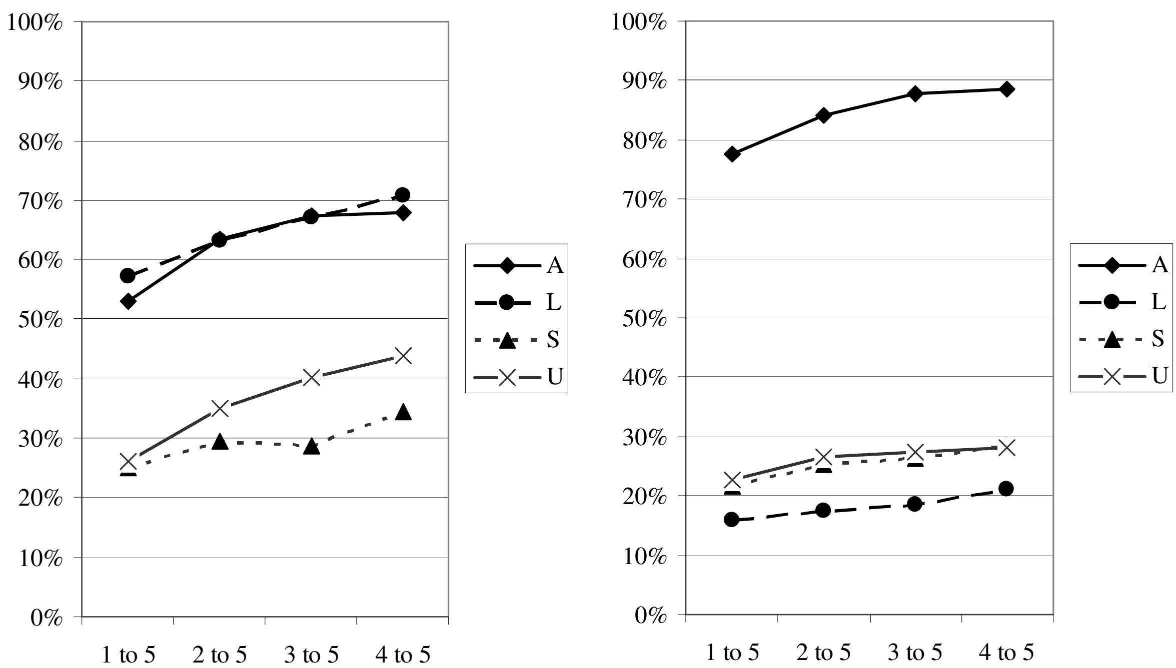


Figure 1: SFP for coarse codes over various samples by school level

Table 3: SFP of the coarse codes over various samples, by school level

Coarse code	Primary				Secondary			
	1 to 5	2 to 5	3 to 5	4 to 5	1 to 5	2 to 5	3 to 5	4 to 5
A	53	64	67	68	78	84	88	88
L	57	63	67	71	16	17	19	21
S	25	30	29	35	22	25	26	28
U	26	35	40	44	23	26	27	28

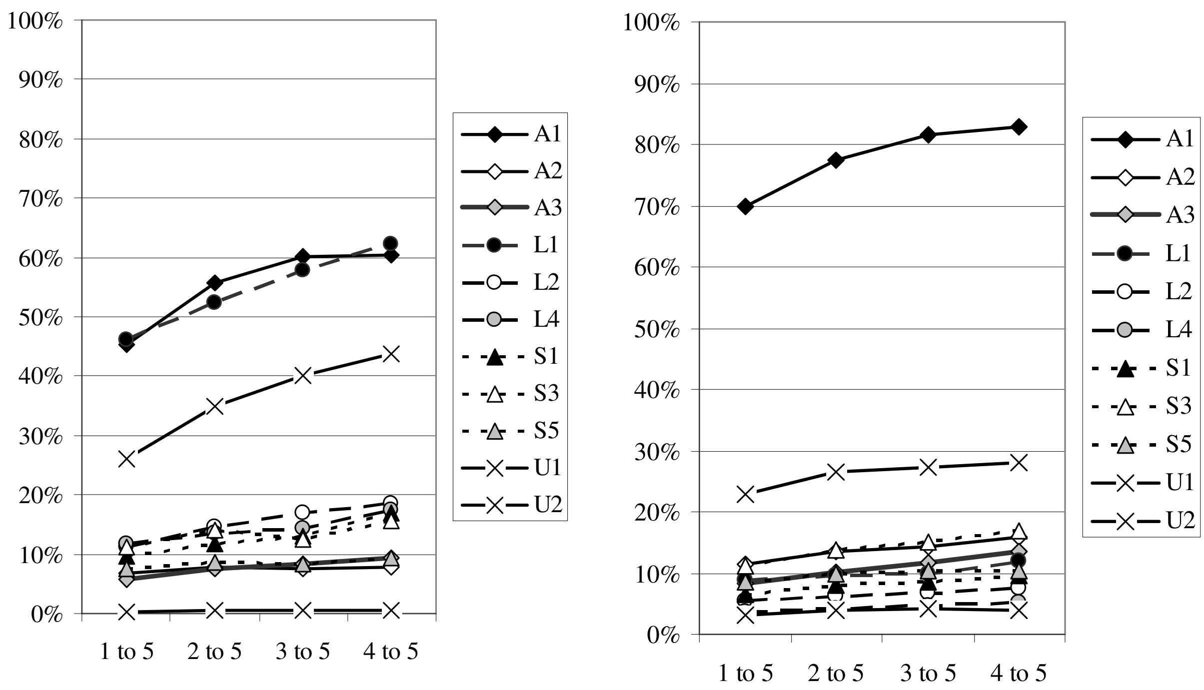


Figure 2: SFP for fine codes over various samples by school level

Table 4: SFP of the fine codes over various samples, by school level

Fine code	Primary				Secondary			
	1 to 5	2 to 5	3 to 5	4 to 5	1 to 5	2 to 5	3 to 5	4 to 5
A1	45	56	60	60	70	77	82	83
A2	7	8	8	8	11	13	14	16
A3	6	7	8	9	8	10	12	13
L1	46	52	58	62	9	10	10	12
L2	11	15	17	19	5	6	7	7
L4	12	14	14	17	4	4	5	5
S1	10	12	13	17	7	8	8	10
S3	11	14	13	16	11	14	15	17
S5	8	9	8	9	9	10	10	10
U1	26	35	40	44	23	26	27	28
U2	0	0	1	1	3	4	4	4

Appendix 5

Appendix 6: Condensed test histories

Table 1: Grades of first and third non-U tests for sample of 1663 students

Grade of first non-U test	Grade of third non-U test						Total	%
	Gr 5	Gr 6	Gr 7	Gr 8	Gr 9	Gr 10		
Grade 4	49	154	19	5			227	*14%*
Grade 5		216	88	9	1		314	*19%*
Grade 6			99	65	8		172	*10%*
Grade 7				364	280	64	708	*43%*
Grade 8					71	105	176	*11%*
Grade 9						66	66	*4%*
Total	49	370	206	443	360	235	1663	*100%*
%	*3%*	*22%*	*12%*	*27%*	*22%*	*14%*	*100%*	

Table 2: Numbers of students from each grade by coarse code of first non-U test

First coarse non-U test	Grade of first non-U test						Total	%
	4	5	6	7	8	9		
A	15	104	91	463	121	63	857	*52%*
L	187	148	43	118	23	2	521	*31%*
S	25	62	38	127	32	1	285	*17%*
Total	227	314	172	708	176	66	1663	*100%*
%	*14%*	*19%*	*10%*	*43%*	*11%*	*4%*	*100%*	

Table 3: Grade distribution of students by coarse code of first non-U test

First coarse non-U test		Grade of first non-U test					
		4	5	6	7	8	9
A	(n=857)	2%	12%	11%	54%	14%	7%
L	(n=521)	36%	28%	8%	23%	4%	0%
S	(n=285)	9%	22%	13%	45%	11%	0%
Overall	(n=1663)	14%	19%	10%	43%	11%	4%

Table 4: Distribution of second and third non-U tests, given first is A (n=857)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	792	2	8	802
L	12	5	2	19
S	20	3	13	36
Total	824	10	23	857

Table 5: Distribution of second and third non-U tests, given first is L (n=521)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	170	11	9	190
L	84	138	26	248
S	34	18	31	83
Total	288	167	66	521

Table 6: Distribution of second and third non-U tests, given first is S (n=285)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	120	1	11	132
L	17	12	7	36
S	48	7	62	117
Total	185	20	80	285

Table 7: Numbers of students from each grade by fine code of first non-U test

First fine non-U test	Grade of first non-U test						Total	%
	4	5	6	7	8	9		
A1	10	94	73	399	101	56	733	*44%*
A2	3	6	16	40	17	4	86	*5%*
A3	2	4	2	24	3	3	38	*2%*
L1	158	113	23	69	11	1	375	*23%*
L2	14	18	7	24	7	0	70	*4%*
L4	15	17	13	25	5	1	76	*5%*
S1	7	20	14	41	6	0	88	*5%*
S3	9	20	12	59	20	1	121	*7%*
S5	9	22	12	27	6	0	76	*5%*
Total	227	314	172	708	176	66	1663	*100%*
%	*14%*	*19%*	*10%*	*43%*	*11%*	*4%*	*100%*	

Table 8: Grade distribution of students by fine code of first non-U test

First fine non-U test		Grade of first non-U test					
		4	5	6	7	8	9
A1	(n=733)	1%	13%	10%	54%	14%	8%
A2	(n=86)	3%	7%	19%	47%	20%	5%
A3	(n=38)	5%	11%	5%	63%	8%	8%
L1	(n=375)	42%	30%	6%	18%	3%	0%
L2	(n=70)	20%	26%	10%	34%	10%	0%
L4	(n=76)	20%	22%	17%	33%	7%	1%
S1	(n=88)	8%	23%	16%	47%	7%	0%
S3	(n=121)	7%	17%	10%	49%	17%	1%
S5	(n=76)	12%	29%	16%	36%	8%	0%
Overall	(n=1663)	14%	19%	10%	43%	11%	4%

Table 9: Distribution of second and third non-U tests, given first is A1 (n=733)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	697	2	5	704
L	7	1	1	9
S	16	2	2	20
Total	720	5	8	733

Table 10: Distribution of second and third non-U tests, given first is A2 (n=86)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	66	0	1	67
L	2	2	1	5
S	3	1	10	14
Total	71	3	12	86

Table 11: Distribution of second and third non-U tests, given first is A3 (n=38)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	29	0	2	31
L	3	2	0	5
S	1	0	1	2
Total	33	2	3	38

Table 12: Distribution of second and third non-U tests, given first is L1 (n=375)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	96	8	5	109
L	68	118	21	207
S	21	14	24	59
Total	185	140	50	375

Table 13: Distribution of second and third non-U tests, given first is L2 (n=70)

Coarse code on second non-U test	Coarse code on third non-U test			Total
	A	L	S	
A	39	0	1	40
L	4	9	4	17
S	7	1	5	13
Total	50	10	10	70

Table 14: Distribution of second and third non-U tests, given first is L4 (n=76)

Coarse code on second non-U test	Coarse code on third non-U test A	L	S	Total
A	35	3	3	41
L	12	11	1	24
S	6	3	2	11
Total	53	17	6	76

Table 15: Distribution of second and third non-U tests, given first is S1 (n=88)

Coarse code on second non-U test	Coarse code on third non-U test A	L	S	Total
A	43	0	4	47
L	5	1	0	6
S	17	2	16	35
Total	65	3	20	88

Table 16: Distribution of second and third non-U tests, given first is S3 (n=121)

Coarse code on second non-U test	Coarse code on third non-U test A	L	S	Total
A	44	0	4	48
L	8	8	4	20
S	19	2	32	53
Total	71	10	40	121

Table 17: Distribution of second and third non-U tests, given first is S5 (n=76)

Coarse code on second non-U test	Coarse code on third non-U test A	L	S	Total
A	33	1	3	37
L	4	3	3	10
S	12	3	14	29
Total	49	7	20	76

Appendix 7: Transitions between consecutive tests

Table 1: Distribution (%) of next coarse code given current fine code

Fine code on current test T_i		Coarse code on next test T_{i+1} A	L	S	U
A1	(n=3279)	93	1	2	4
A2	(n=280)	78	4	10	10
A3	(n=187)	68	6	4	22
L1	(n=853)	21	52	11	17
L2	(n=227)	42	23	14	20
L4	(n=177)	35	32	14	19
S1	(n=245)	46	9	32	12
S3	(n=385)	27	9	46	18
S5	(n=217)	40	8	30	23
U1	(n=757)	46	10	15	29
U2	(n=51)	53	2	6	40
Overall	(n=6658)	66	12	10	12

Table 2: Distribution (%) of previous coarse code given current fine code

Fine code on Current test T_i		Coarse code on previous test T_{i-1} A	L	S	U
A1	(n=3952)	80	6	6	8
A2	(n=258)	59	13	15	14
A3	(n=218)	40	23	16	21
L1	(n=464)	3	81	9	7
L2	(n=180)	16	59	12	13
L4	(n=106)	5	63	12	20
S1	(n=186)	15	23	42	20
S3	(n=331)	10	19	56	16
S5	(n=147)	19	27	39	16
U1	(n=748)	23	30	18	29
U2	(n=68)	50	3	19	28
Overall	(n=6658)	56	19	13	12
non-A1	*(n=2706)*	*21*	*37*	*23*	*19*

Table 3: Comparison of arrival rate and departure rate for each code

Previous test T_{i-1} Retest	Arrivals	**Current test T_i**	Next test T_{i+1} Retest	Departures	Departures-Arrivals
77	23	**A**	91	9	-14
73	27	**L**	44	56	+29
48	52	**S**	38	62	+10
29	71	**U**	29	71	0
74	26	**A1**	89	11	-15
21	79	**A2**	19	81	+2
11	89	**A3**	12	88	-2
69	31	**L1**	38	62	+32
14	86	**L2**	11	89	+3
9	91	**L4**	6	94	+4
17	83	**S1**	13	87	+4
38	62	**S3**	33	67	+5
10	90	**S5**	7	93	+3
28	72	**U1**	28	72	0
18	82	**U2**	24	76	-6

Table 4: Distribution (%) of next test given coarse code of current test (primary)

Current test		Next test A	L	S	U
A	(n=801)	90	3	2	5
L	(n=867)	23	48	12	17
S	(n=311)	40	15	30	14
U	(n=322)	44	16	13	27
Non-A	*(n=1500)*	*31*	*34*	*16*	*18*

Table 5: Distribution (%) of next test given coarse code of current test (secondary)

Current test		Next test A	L	S	U
A	(n=2945)	91	1	2	5
L	(n=390)	34	35	9	21
S	(n=536)	33	5	42	19
U	(n=486)	49	5	15	31
Non-A	*(n=1412)*	*39*	*13*	*24*	*24*

Table 6: Distribution (%) of next test given fine code of current test (primary)

Current test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
A1	(n=700)	90	2	1	0	1	0	1	1	0	4	0
A2	(n=57)	67	7	2	4	2	4	2	5	0	7	2
A3	(n=44)	41	2	16	11	2	0	5	0	2	20	0
L1	(n=633)	13	2	3	39	8	6	4	5	3	17	0
L2	(n=122)	31	7	2	13	13	2	7	7	3	14	1
L4	(n=112)	26	1	4	20	11	6	6	4	4	19	0
S1	(n=107)	44	4	7	9	2	3	15	6	3	7	1
S3	(n=129)	29	2	1	12	2	2	8	26	4	16	0
S5	(n=75)	28	3	3	11	3	4	8	16	4	21	0
U1	(n=321)	36	3	4	8	4	4	6	5	2	27	0
U2	(n=1)	0	0	0	100	0	0	0	0	0	0	0
Overall	(n=2301)	46	3	3	16	5	3	4	5	2	14	0
Non-A1	*(n=1601)*	*27*	*3*	*4*	*22*	*6*	*4*	*6*	*7*	*2*	*18*	*0*

Table 7: Distribution (%) of next test given fine code of current test (secondary)

Current test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
A1	(n=2579)	89	3	2	0	0	0	0	0	1	3	1
A2	(n=223)	49	22	7	0	2	0	3	5	3	8	2
A3	(n=143)	52	7	11	1	2	1	0	1	1	19	3
L1	(n=220)	20	2	5	33	7	6	0	4	3	18	0
L2	(n=105)	33	5	7	7	10	1	2	6	3	28	0
L4	(n=65)	32	3	6	11	8	5	0	8	8	20	0
S1	(n=138)	25	10	6	1	4	1	12	14	12	14	1
S3	(n=256)	19	3	3	2	3	1	7	37	7	16	2
S5	(n=142)	30	6	7	1	1	1	9	13	8	21	3
U1	(n=436)	37	5	6	1	3	2	4	8	3	28	2
U2	(n=50)	44	2	8	0	0	0	0	4	2	16	24
Overall	(n=4357)	66	4	3	2	2	1	2	5	2	10	1
Non-A1	*(n=1778)*	*33*	*7*	*6*	*6*	*4*	*2*	*4*	*11*	*5*	*20*	*2*

Table 8: Distribution (%) of the next test after A test by current grade

Grade of A test		Next test A	L	S	U
Grade 4/5	(n=233)	81	6	4	9
Grade 6	(n=568)	93	2	2	4
Grade 7	(n=1202)	91	1	3	5
Grade 8	(n=1081)	92	1	2	5
Grade 9	(n=525)	90	1	2	7
Grade 10	(n=137)	94	0	1	5
Overall	**(n=3746)**	**91**	**1**	**2**	**5**

Table 9: Distribution (%) of the next test after A test by school group & school level

School Group		Next test A	L	S	U
Primary	**(n=801)**	**90**	**3**	**2**	**5**
SGA	(n=28)	79	11	0	11
SGB	(n=70)	89	0	9	3
SGC	(n=211)	91	3	1	5
SGD	(n=301)	89	2	3	7
SGE	(n=191)	92	4	1	3
Secondary	**(n=2945)**	**91**	**1**	**2**	**5**
SGA	(n=359)	86	2	4	9
SGB	(n=392)	94	1	2	3
SGC	(n=601)	90	1	3	5
SGD	(n=402)	92	0	2	5
SGE	(n=460)	97	0	1	2
SGF	(n=731)	90	1	2	7
Secondary-Primary		**+2**	**-2**	**0**	**0**

Table 10: Distribution (%) of the next test after A1 test by current grade

Grade of A1 test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=188)	84	3	0	1	2	0	2	2	0	7	0
Grade 6	(n=512)	92	2	1	0	1	0	1	0	0	3	0
Grade 7	(n=1057)	90	3	1	0	0	0	0	1	1	3	1
Grade 8	(n=953)	89	3	2	0	0	0	1	0	1	3	1
Grade 9	(n=449)	88	4	2	0	0	0	0	0	0	4	1
Grade 10	(n=120)	93	3	2	0	0	0	0	1	0	1	1
Overall	**(n=3279)**	**89**	**3**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**3**	**1**

Table 11: Distribution (%) of the next test after A1 test by school group & school level

School Group		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=700)**	**90**	**2**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**4**	**0**
SGA	(n=21)	90	5	0	0	0	0	0	0	0	5	0
SGB	(n=61)	87	3	2	0	0	0	2	5	0	2	0
SGC	(n=171)	91	3	1	1	1	0	1	0	0	2	1
SGD	(n=264)	88	2	1	0	2	0	2	0	0	6	0
SGE	(n=183)	92	1	1	1	2	0	0	1	0	3	0
Secondary	**(n=2579)**	**89**	**3**	**2**	**0**	**0**	**0**	**0**	**0**	**1**	**3**	**1**
SGA	(n=306)	84	5	1	0	1	0	0	1	1	5	1
SGB	(n=352)	92	2	2	0	0	0	1	1	1	1	1
SGC	(n=527)	90	2	1	0	0	0	1	1	1	4	0
SGD	(n=354)	89	4	1	0	0	0	0	0	1	4	0
SGE	(n=432)	96	1	1	0	0	0	0	0	0	1	1
SGF	(n=608)	86	3	3	0	0	0	0	0	1	4	1
Secondary-Primary		**0**	**+1**	**+1**	**0**	**-1**	**0**	**-1**	**0**	**+1**	**-1**	**+1**

Table 12: Distribution (%) of the next test after A2 test by current grade

Grade of A2 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=24)	67	4	0	4	0	8	0	4	0	8	4
Grade 6	(n=33)	67	9	3	3	3	0	3	6	0	6	0
Grade 7	(n=88)	55	16	2	0	3	0	2	8	6	7	1
Grade 8	(n=79)	54	24	6	0	1	0	3	3	0	5	4
Grade 9/10	(n=56)	32	29	16	0	0	0	4	4	2	13	2
Overall	**(n=280)**	**53**	**19**	**6**	**1**	**2**	**1**	**3**	**5**	**2**	**8**	**2**

Table 13: Distribution (%) of the next test after A2 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=57)**	**67**	**7**	**2**	**4**	**2**	**4**	**2**	**5**	**0**	**7**	**2**
SGA	(n=4)	50	0	0	0	0	25	0	0	0	25	0
SGB	(n=7)	57	0	0	0	0	0	0	29	0	0	14
SGC	(n=19)	74	5	0	5	0	5	0	0	0	11	0
SGD	(n=22)	73	14	5	0	0	0	5	0	0	5	0
SGE	(n=5)	40	0	0	20	20	0	0	20	0	0	0
Secondary	**(n=223)**	**49**	**22**	**7**	**0**	**2**	**0**	**3**	**5**	**3**	**8**	**2**
SGA	(n=33)	52	9	6	0	0	0	0	12	3	15	3
SGB	(n=20)	65	10	0	0	5	0	0	10	0	5	5
SGC	(n=41)	37	27	5	0	7	0	5	2	7	7	2
SGD	(n=29)	66	21	3	0	0	0	3	3	3	0	0
SGE	(n=20)	65	10	5	0	0	0	0	10	0	5	5
SGF	(n=80)	40	31	13	0	0	0	4	1	1	9	1
Secondary-Primary		**-18**	**+15**	**+5**	**-4**	**0**	**-4**	**+1**	**0**	**+3**	**+1**	**0**

Table 14: Distribution (%) of the next test after A3 test by current grade

Grade of A3 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=21)	33	0	10	19	5	0	5	0	5	24	0
Grade 6	(n=23)	48	4	22	4	0	0	4	0	0	17	0
Grade 7	(n=57)	54	7	14	2	2	2	0	0	4	14	2
Grade 8	(n=49)	49	6	14	0	0	2	0	4	0	20	4
Grade 9/10	(n=37)	54	8	3	0	5	0	0	0	0	24	5
Overall	**(n=187)**	**50**	**6**	**12**	**3**	**2**	**1**	**1**	**1**	**2**	**19**	**3**

Table 15: Distribution (%) of the next test after A3 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=44)**	**41**	**2**	**16**	**11**	**2**	**0**	**5**	**0**	**2**	**20**	**0**
SGA	(n=3)	0	0	0	67	0	0	0	0	0	33	0
SGB	(n=2)	100	0	0	0	0	0	0	0	0	0	0
SGC	(n=21)	38	0	29	10	0	0	5	0	5	14	0
SGD	(n=15)	47	0	7	0	7	0	7	0	0	33	0
SGE	(n=3)	33	33	0	33	0	0	0	0	0	0	0
Secondary	**(n=143)**	**52**	**7**	**11**	**1**	**2**	**1**	**0**	**1**	**1**	**19**	**3**
SGA	(n=20)	40	5	5	0	10	5	0	0	0	30	5
SGB	(n=20)	55	5	25	0	0	0	0	0	5	10	0
SGC	(n=33)	58	0	15	0	0	3	0	3	3	18	0
SGD	(n=19)	42	5	21	5	0	0	0	5	0	11	11
SGE	(n=8)	63	25	0	0	0	0	0	0	0	13	0
SGF	(n=43)	56	12	2	0	2	0	0	0	0	23	5
Secondary-Primary		**+12**	**+5**	**-5**	**-11**	**0**	**+1**	**-5**	**+1**	**-1**	**-2**	**+3**

Table 16: Distribution (%) of the next test after L test by current grade

Grade of L test		Next test A	L	S	U
Grade 4	**(n=210)**	18	53	16	13
Grade 5	(n=393)	18	51	12	19
Grade 6	(n=264)	35	38	10	17
Grade 7	(n=243)	35	34	11	20
Grade 8	(n=115)	35	32	8	25
Grade 9/10	(n=32)	28	50	6	16
Overall	**(n=1257)**	**27**	**44**	**12**	**18**

Table 17: Distribution (%) of the next test after L test by school group & school level

School Group		Next test A	L	S	U
Primary	**(n=867)**	**23**	**48**	**12**	**17**
SGA	(n=204)	16	59	7	18
SGB	(n=62)	24	29	34	13
SGC	(n=380)	22	55	9	14
SGD	(n=186)	31	30	19	20
SGE	(n=35)	40	23	9	29
Secondary	**(n=390)**	**34**	**35**	**9**	**21**
SGA	(n=113)	24	44	10	22
SGB	(n=58)	34	41	7	17
SGC	(n=106)	29	37	8	26
SGD	(n=39)	33	36	13	18
SGE	(n=**11)**	55	18	9	18
SGF	(n=63)	59	11	13	17
Secondary-Primary		**+11**	**-13**	**-3**	**+5**

Table 18: Distribution (%) of the next test after L1 test by current grade

Grade of L1 test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4	(n=179)	11	3	4	36	8	7	5	7	3	15	0
Grade 5	(n=290)	9	2	4	46	6	6	3	4	2	18	0
Grade 6	(n=164)	23	2	2	32	9	6	4	4	2	16	0
Grade 7	(n=153)	22	1	5	31	8	6	1	4	3	18	1
Grade 8-10	(n=67)	16	3	4	39	6	7	0	3	3	18	0
Overall	**(n=853)**	**15**	**2**	**4**	**38**	**8**	**6**	**3**	**5**	**3**	**17**	**0**

Table 19: Distribution (%) of the next test after L1 test by school group & school level

School Group		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=633)**	**13**	**2**	**3**	**39**	**8**	**6**	**4**	**5**	**3**	**17**	**0**
SGA	(n=176)	7	3	2	53	6	4	1	5	1	17	0
SGB	(n=48)	17	0	6	23	10	0	4	19	8	13	0
SGC	(n=267)	13	1	4	43	9	8	3	2	1	15	0
SGD	(n=124)	19	4	5	20	7	6	9	7	5	19	0
SGE	(n=18)	22	0	0	22	0	17	6	0	0	33	0
Secondary	**(n=220)**	**20**	**2**	**5**	**33**	**7**	**6**	**0**	**4**	**3**	**18**	**0**
SGA	(n=75)	12	1	3	37	12	8	0	4	4	19	0
SGB	(n=28)	21	4	7	39	4	11	0	0	0	14	0
SGC	(n=61)	15	2	5	36	7	7	0	5	2	21	2
SGD	(n=24)	21	4	8	25	8	4	0	4	8	17	0
SGE	(n=4)	50	0	25	0	0	0	0	0	0	25	0
SGF	(n=28)	50	0	4	21	0	0	4	4	4	14	0
Secondary-Primary		**+7**	**0**	**+2**	**-6**	**0**	**0**	**-3**	**-1**	**+1**	**+1**	**0**

Table 20: Distribution (%) of the next test after L2 test by current grade

Grade of L2 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=65)	18	6	2	22	15	3	9	11	3	9	2
Grade 6	(n=57)	46	7	2	4	11	2	5	2	4	19	0
Grade 7	(n=48)	29	10	8	6	8	0	2	8	4	23	0
Grade 8-10	(n=57)	37	0	5	7	11	2	2	4	2	32	0
Overall	**(n=227)**	**32**	**6**	**4**	**10**	**11**	**2**	**5**	**6**	**3**	**20**	**0**

Table 21: Distribution (%) of the next test after L2 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=122)**	**31**	**7**	**2**	**13**	**13**	**2**	**7**	**7**	**3**	**14**	**1**
SGA	(n=14)	21	7	0	21	14	0	0	0	7	29	0
SGB	(n=7)	29	0	0	0	0	0	14	43	0	14	0
SGC	(n=57)	28	5	2	18	19	5	7	4	2	9	2
SGD	(n=35)	34	11	0	9	9	0	11	6	6	14	0
SGE	(n=9)	56	0	11	0	0	0	0	11	0	22	0
Secondary	**(n=105)**	**33**	**5**	**7**	**7**	**10**	**1**	**2**	**6**	**3**	**28**	**0**
SGA	(n=22)	45	0	5	5	14	0	0	9	0	23	0
SGB	(n=22)	41	5	0	9	14	5	5	0	5	18	0
SGC	(n=27)	15	7	11	11	4	0	0	4	4	44	0
SGD	(n=11)	18	9	9	9	18	0	0	9	0	27	0
SGE	(n=4)	50	0	0	0	0	0	0	0	25	25	0
SGF	(n=19)	42	5	11	0	5	0	5	11	0	21	0
Secondary-Primary		**+2**	**-2**	**+5**	**-6**	**-4**	**-2**	**-5**	**-1**	**0**	**+14**	**-1**

Table 22: Distribution (%) of the next test after L4 test by current grade

Grade of L4 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=69)	23	0	3	22	9	7	6	6	4	20	0
Grade 6	(n=43)	30	2	7	16	14	5	7	0	2	16	0
Grade 7	(n=42)	36	0	7	10	5	5	0	12	5	21	0
Grade 8-10	(n=23)	26	9	4	13	13	4	0	0	13	17	0
Overall	**(n=177)**	**28**	**2**	**5**	**16**	**10**	**6**	**4**	**5**	**5**	**19**	**0**

Table 23: Distribution (%) of the next test after L4 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=112)**	**26**	**1**	**4**	**20**	**11**	**6**	**6**	**4**	**4**	**19**	**0**
SGA	(n=14)	36	0	7	21	14	0	7	0	0	14	0
SGB	(n=7)	29	0	0	14	14	0	0	29	0	14	0
SGC	(n=56)	23	0	5	21	14	9	5	4	7	11	0
SGD	(n=27)	22	0	4	19	4	7	7	0	0	37	0
SGE	(n=8)	38	13	0	13	0	0	13	0	0	25	0
Secondary	**(n=65)**	**32**	**3**	**6**	**11**	**8**	**5**	**0**	**8**	**8**	**20**	**0**
SGA	(n=16)	25	0	0	6	6	6	0	6	13	38	0
SGB	(n=8)	13	0	0	25	13	0	0	13	13	25	0
SGC	(n=18)	39	0	11	6	11	11	0	6	6	11	0
SGD	(n=4)	25	0	0	50	0	0	0	25	0	0	0
SGE	(n=3)	33	0	0	33	33	0	0	0	0	0	0
SGF	(n=16)	44	13	13	0	0	0	0	6	6	19	0
Secondary-Primary		**+6**	**+2**	**+2**	**-9**	**-3**	**-2**	**-6**	**+4**	**+4**	**+1**	**0**

Table 24: Distribution (%) of the next test after S test by current grade

Grade of S test		Next test A	L	S	U
Grade 4	**(n=27)**	26	19	33	22
Grade 5	(n=134)	36	16	32	16
Grade 6	(n=150)	**46**	15	**28**	**11**
Grade 7	(n=252)	35	6	40	20
Grade 8	(n=195)	**27**	7	**49**	18
Grade 9/10	(n=89)	44	1	34	21
Overall	**(n=847)**	**36**	**9**	**38**	**18**

Table 25: Distribution (%) of the next test after S test by school group & school level

School Group		Next test A	L	S	U
Primary	**(n=311)**	**40**	**15**	**30**	**14**
SGA	(n=28)	36	18	29	18
SGB	(n=56)	43	7	36	14
SGC	(n=108)	38	15	36	11
SGD	(n=97)	40	23	23	14
SGE	(n=22)	45	5	23	27
Secondary	**(n=536)**	**33**	**5**	**42**	**19**
SGA	(n=96)	35	7	33	24
SGB	(n=98)	44	3	36	17
SGC	(n=82)	20	10	50	21
SGD	(n=73)	10	7	66	18
SGE	(n=48)	35	0	50	15
SGF	(n=139)	45	4	32	19
Secondary-Primary		**-6**	**-10**	**+12**	**+5**

Table 26: Distribution (%) of the next test after S1 test by current grade

Grade of S1 test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=50)	36	6	2	14	0	6	16	8	2	10	0
Grade 6	(n=57)	51	2	11	5	4	0	14	4	4	5	2
Grade 7	(n=80)	20	9	5	1	4	1	15	15	11	18	1
Grade 8	(n=35)	29	9	6	0	9	0	9	14	14	9	3
Grade 9/10	(n=23)	35	17	9	0	0	0	4	13	9	13	0
Overall	**(n=245)**	**33**	**7**	**6**	**4**	**3**	**2**	**13**	**11**	**8**	**11**	**1**

Table 27: Distribution (%) of the next test after S1 test by school group & school level

School Group		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=107)**	**44**	**4**	**7**	**9**	**2**	**3**	**15**	**6**	**3**	**7**	**1**
SGA	(n=7)	29	0	0	0	0	14	57	0	0	0	0
SGB	(n=15)	60	7	7	0	0	0	7	7	7	0	7
SGC	(n=41)	46	5	5	7	2	0	22	10	0	2	0
SGD	(n=37)	35	3	8	19	3	5	3	3	5	16	0
SGE	(n=7)	57	0	14	0	0	0	14	0	0	14	0
Secondary	**(n=138)**	**25**	**10**	**6**	**1**	**4**	**1**	**12**	**14**	**12**	**14**	**1**
SGA	(n=24)	38	4	4	4	4	0	8	17	8	13	0
SGB	(n=31)	35	6	6	0	0	0	6	13	13	19	0
SGC	(n=18)	11	11	0	0	6	6	17	17	17	11	6
SGD	(n=19)	11	0	5	0	16	0	26	21	11	11	0
SGE	(n=14)	21	29	0	0	0	0	14	7	14	14	0
SGF	(n=32)	22	16	13	0	3	0	6	13	9	16	3
Secondary-Primary		**-19**	**+6**	**-1**	**-9**	**+2**	**-2**	**-3**	**+9**	**+9**	**+7**	**+1**

Table 28: Distribution (%) of the next test after S3 test by current grade

Grade of S3 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=70)	30	1	0	10	0	1	7	24	4	21	0
Grade 6	(n=59)	29	2	2	15	3	2	8	27	3	8	0
Grade 7	(n=107)	28	3	1	4	3	0	5	35	6	15	2
Grade 8	(n=116)	**13**	2	3	1	4	3	8	41	8	16	2
Grade 9/10	(n=33)	9	9	6	0	0	0	9	30	9	21	6
Overall	**(n=385)**	**22**	**3**	**2**	**5**	**3**	**1**	**7**	**33**	**6**	**16**	**2**

Table 29: Distribution (%) of the next test after S3 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=129)**	**29**	**2**	**1**	**12**	**2**	**2**	**8**	**26**	**4**	**16**	**0**
SGA	(n=12)	25	0	0	8	0	8	8	0	8	42	0
SGB	(n=27)	33	0	4	7	0	0	15	33	4	4	0
SGC	(n=42)	21	2	0	12	2	2	5	33	5	17	0
SGD	(n=40)	38	3	0	18	3	0	8	20	3	10	0
SGE	(n=8)	25	0	0	13	0	0	0	25	0	38	0
Secondary	**(n=256)**	**19**	**3**	**3**	**2**	**3**	**1**	**7**	**37**	**7**	**16**	**2**
SGA	(n=37)	19	0	8	0	5	3	5	24	11	24	0
SGB	(n=40)	25	8	0	3	3	3	8	25	13	15	0
SGC	(n=44)	18	2	0	7	2	0	9	43	2	11	5
SGD	(n=42)	5	0	0	2	2	0	5	64	5	17	0
SGE	(n=22)	5	5	5	0	0	0	5	59	9	14	0
SGF	(n=71)	28	4	4	0	4	1	7	23	6	17	6
Secondary-Primary		**-11**	**+2**	**+2**	**-10**	**+2**	**0**	**-1**	**+11**	**+3**	**+1**	**+2**

Table 30: Distribution (%) of the next test after S5 test by current grade

Grade of S5 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4/5	(n=41)	20	2	5	15	2	2	12	20	2	20	0
Grade 6	(n=34)	38	3	0	6	3	6	3	12	6	24	0
Grade 7	(n=65)	31	5	6	2	2	0	9	18	2	23	3
Grade 8	(n=44)	23	7	7	0	2	0	14	11	**14**	20	2
Grade 9/10	(n=33)	36	6	9	0	0	3	3	6	15	18	3
Overall	**(n=217)**	**29**	**5**	**6**	**4**	**2**	**2**	**9**	**14**	**7**	**21**	**2**

Table 31: Distribution (%) of the next test after S5 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=75)**	**28**	**3**	**3**	**11**	**3**	**4**	**8**	**16**	**4**	**21**	**0**
SGA	(n=9)	44	0	11	11	0	11	0	11	11	0	0
SGB	(n=14)	21	0	0	7	0	7	14	7	0	43	0
SGC	(n=25)	24	4	4	12	8	0	4	20	8	16	0
SGD	(n=20)	25	5	0	15	0	5	10	20	0	20	0
SGE	(n=7)	43	0	0	0	0	0	14	14	0	29	0
Secondary	**(n=142)**	**30**	**6**	**7**	**1**	**1**	**1**	**9**	**13**	**8**	**21**	**3**
SGA	(n=35)	26	3	9	3	0	3	3	11	11	29	3
SGB	(n=27)	37	11	7	0	0	0	11	11	4	19	0
SGC	(n=20)	10	0	5	0	10	0	15	15	10	25	10
SGD	(n=12)	17	0	0	0	0	0	25	25	0	33	0
SGE	(n=12)	25	17	17	0	0	0	8	17	0	17	0
SGF	(n=36)	44	6	6	0	0	0	6	11	14	11	3
Secondary-Primary		**+2**	**+3**	**+4**	**-10**	**-1**	**-3**	**+1**	**-3**	**+4**	**0**	**+3**

Table 32: Distribution (%) of the next test after U test by current grade

Grade of U test		Next test A	L	S	U
Grade 4	(n=33)	12	52	30	6
Grade 5	(n=142)	44	12	15	29
Grade 6	(n=147)	50	13	7	30
Grade 7	(n=195)	48	7	15	30
Grade 8	(n=164)	51	4	13	32
Grade 9	(n=100)	49	4	20	27
Grade 10	(n=27)	48	0	7	44
Overall	**(n=808)**	**47**	**10**	**14**	**29**

Table 33: Distribution (%) of the next test after U test by school group & school level

School Group		Next test A	L	S	U
Primary	**(n=322)**	**44**	**16**	**13**	**27**
SGA	(n=47)	38	21	6	34
SGB	(n=21)	57	10	14	19
SGC	(n=115)	33	19	16	32
SGD	(n=94)	53	16	11	20
SGE	(n=45)	51	9	16	24
Secondary	**(n=486)**	**49**	**5**	**15**	**31**
SGA	(n=87)	48	6	20	26
SGB	(n=67)	49	7	16	27
SGC	(n=105)	43	9	11	37
SGD	(n=65)	42	6	18	34
SGE	(n=23)	70	0	4	26
SGF	(n=139)	55	1	14	29
Secondary-Primary		**+6**	**-11**	**+2**	**+4**

Table 34: Distribution (%) of the next test after U1 test by current grade

Grade of U1 test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 4	(n=33)	9	3	0	27	12	12	9	15	6	6	0
Grade 5	(n=141)	39	4	2	5	2	4	8	5	2	29	0
Grade 6	(n=147)	40	3	7	7	3	2	3	2	1	30	0
Grade 7	(n=188)	36	5	7	1	4	3	5	7	3	28	1
Grade 8	(n=140)	38	5	6	1	2	1	2	9	3	29	2
Grade 9	(n=88)	41	6	5	2	1	1	7	9	6	20	2
Grade 10	(n=20)	30	5	10	0	0	0	0	5	0	50	0
Overall	**(n=757)**	**37**	**4**	**5**	**4**	**3**	**3**	**5**	**7**	**3**	**28**	**1**

Table 35: Distribution (%) of the next test after U1 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=321)**	**36**	**3**	**4**	**8**	**4**	**4**	**6**	**5**	**2**	**27**	**0**
SGA	(n=46)	24	4	11	13	2	4	0	2	4	35	0
SGB	(n=21)	38	10	10	5	0	5	0	10	5	19	0
SGC	(n=115)	24	4	4	11	3	5	9	5	2	32	0
SGD	(n=94)	51	1	1	7	5	3	7	3	0	20	0
SGE	(n=45)	49	2	0	0	7	2	4	7	4	24	0
Secondary	**(n=436)**	**37**	**5**	**6**	**1**	**3**	**2**	**4**	**8**	**3**	**28**	**2**
SGA	(n=78)	40	3	5	1	3	3	1	17	4	24	0
SGB	(n=57)	39	0	11	5	4	0	9	7	4	19	4
SGC	(n=99)	29	9	3	2	4	3	2	5	5	35	2
SGD	(n=61)	31	3	8	0	3	3	10	8	2	28	3
SGE	(n=20)	45	10	15	0	0	0	5	0	0	25	0
SGF	(n=121)	43	7	6	0	1	1	3	7	3	29	1
Secondary-Primary		**+1**	**+2**	**+2**	**-7**	**-1**	**-2**	**-2**	**+3**	**+1**	**+1**	**+2**

Table 36: Distribution (%) of the next test after U2 test by current grade

Grade of U2 test		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Grade 5-8	(n=32)	50	0	9	3	0	0	0	0	3	19	16
Grade 9/10	(n=19)	32	5	5	0	0	0	0	11	0	11	37
Overall	**(n=51)**	**43**	**2**	**8**	**2**	**0**	**0**	**0**	**4**	**2**	**16**	**24**

Table 37: Distribution (%) of the next test after U2 test by school group & school level

School Group		Next test										
		A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Primary	**(n=1)**	**0**	**0**	**0**	**100**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Secondary	**(n=50)**	**44**	**2**	**8**	**0**	**0**	**0**	**0**	**4**	**2**	**16**	**24**
SGA	(n=9)	44	0	11	0	0	0	0	0	0	11	33
SGB	(n=10)	30	0	20	0	0	0	0	0	0	30	20
SGC	(n=6)	67	0	0	0	0	0	0	0	0	17	17
SGD	(n=4)	25	0	0	0	0	0	0	0	0	50	25
SGE	(n=3)	67	0	0	0	0	0	0	0	0	0	33
SGF	(n=18)	44	6	6	0	0	0	0	11	6	6	22

Table 38: Distribution (%) of the next test after non-A test by current grade

Grade of non-A test		Next test			
		A	L	S	U
Gr 4	(n=270)	18	49	19	13
Gr 5	(n=669)	27	36	17	20
Gr 6	(n=561)	42	25	14	19
Gr 7	(n=690)	39	16	22	23
Gr 8	(n=474)	37	12	26	24
Gr 9	(n=199)	44	9	25	23
Gr 10	(n=49)	45	8	10	37
Overall	**(n=2912)**	**35**	**24**	**20**	**21**
Cf all students	*(n=6658)*	*67*	*11*	*10*	*12*
Primary	(n=1500)	31	34	16	18
Secondary	(n=1412)	39	13	24	24
Secondary-Primary		**+8**	**-21**	**+7**	**+5**

Table 39: Distribution (%) of the next test after non-A1 test by current grade

Grade of Non-A1 test		Next test A1	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2
Gr4	(n=275)	12	3	3	32	8	8	6	9	3	13	0
Gr5	(n=709)	24	3	3	25	5	4	6	7	2	20	0
Gr6	(n=617)	37	3	5	14	6	3	6	6	2	18	0
Gr7	(n=835)	33	6	6	8	4	2	4	11	4	19	2
Gr8	(n=602)	33	7	6	4	4	2	4	13	5	20	2
Gr9	(n=275)	33	12	7	4	2	1	5	9	5	17	4
Gr10	(n=66)	38	5	9	5	2	0	0	5	3	29	6
Total	(n=3379)	30	5	5	14	5	3	5	9	4	19	1
Cf all students	*(n=6658)*	*59*	*4*	*3*	*7*	*3*	*2*	*3*	*5*	*2*	*11*	*1*
Primary	(n=1601)	27	3	4	22	6	4	6	7	3	18	0
Secondary	(n=1778)	33	7	6	6	4	2	4	11	5	20	2
Secondary-Primary		**+7**	**+4**	**+2**	**-17**	**-2**	**-3**	**-2**	**+4**	**+2**	**+2**	**+2**

Appendix 8: Sequences and occurrences of codes

Tables 1 to 15 consider each of the codes in turn and compare the total number of occurrences of the given code within students' test histories against the longest sequence. Off-diagonal entries indicate students with scattered occurrences of that code. Tables 16 to 30 contain details of the number of students in each school group with sequences of particular lengths. The number of students involved in each code is used to calculate the Student Focussed Prevalence (SFP); see last column of each table. Tables 31 to 34 contain the sample sizes for the calculations in various figures in Chapter 6.

Table 1: Comparison of longest sequence and total number of occurrences of A

Longest A sequence	Total number of A's in the test history							Total
	1	2	3	4	5	6	7	
A	808	61	1					870
AA		442	44	4				490
AAA			422	21	4			447
AAAA				310	8			318
AAAAA					196	2		198
AAAAAA						48		48
AAAAAAA							5	5
Total	808	503	467	335	208	50	5	2376

Table 2: Comparison of longest sequence and total number of occurrences of A1

Longest A1 sequence	Total number of A1's in the test history							Total
	1	2	3	4	5	6	7	
A1	749	66	1					816
A1A1		400	41	5				446
A1A1A1			372	22	4			398
A1A1A1A1				266	6			272
A1A1A1A1A1					162	1		163
A1A1A1A1A1A1						41		41
A1A1A1A1A1A1A1							1	1
Total	749	466	414	293	172	42	1	2137

Table 3: Comparison of longest sequence and total number of occurrences of A2

Longest A2 sequence	Total number of A2's in the test history				Total
	1	2	3	4	
A2	300	11			311
A2A2		23	4		27
A2A2A2			10		10
A2A2A2A2				2	2
Total	300	34	14	2	350

Table 4: Comparison of longest sequence and total number of occurrences of A3

Longest A3 sequence	Total number of A3's in the test history				Total
	1	2	3	4	
A3	245	5			250
A3A3		12	2		14
A3A3A3			3		3
A3A3A3A3				1	1
Total	245	17	5	1	268

Table 5: Comparison of longest sequence and total number of occurrences of L

Longest L sequence	Total number of L's in the test history						Total
	1	2	3	4	5	6	
L	604	40	2*				646
LL		166	22	2			190
LLL			69	4			73
LLLL				39	3		42
LLLLL					20		20
LLLLLL						1	1
Total	604	206	93	45	23	1	972

* [L1,S1,L1,S1,L1,U1,S1] and [L1,A3,L2,A1,L2,A1]

Table 6: Comparison of longest sequence and total number of occurrences of L1

Longest L1 sequence	Total number of L1's in the test history					Total
	1	2	3	4	5	
L1	455	29	1*			485
L1L1		120	12	1		133
L1L1L1			45	3		48
L1L1L1L1				17	2	19
L1L1L1L1L1					9	9
Total	455	149	58	21	11	694

Table 7: Comparison of longest sequence and total number of occurrences of L2

Longest L2 sequence	Total number of L2's in the test history				Total
	1	2	3	4	
L2	229	8			237
L2L2		20	1	1	22
L2L2L2			2		2
Total	229	28	3	1	261

Table 8: Comparison of longest sequence and total number of occurrences of L4

Longest L4 sequence	Total number of L4's in the test history		Total
	1	2	
L4	204	4	208
L4L4		10	10
Total	204	14	218

Table 9: Comparison of longest sequence and total number of occurrences of S

Longest S sequence	Total number of S's in the test history						Total
	1	2	3	4	5	6	
S	519	33	3				555
SS		112	15	5			132
SSS			44	2		1*	47
SSSS				22	1		23
SSSSS					3		3
SSSSSS						1**	1
Total	519	145	62	29	4	2	761

* [S4,S3,S3,A1,S3,S3,S3]
** [S2,S3,S3,S3,S3,S3]

Table 10: Comparison of longest sequence and total number of occurrences of S1

Longest S1 sequence	Total number of S1's in the test history			Total
	1	2	3	
S1	223	13		236
S1S1		21	5	26
S1S1S1			3	3
Total	223	34	8	265

Table 11: Comparison of longest sequence and total number of occurrences of S3

Longest S3 sequence	Total number of S3's in the test history					Total
	1	2	3	4	5	
S3	279	13				292
S3S3		46	8	2		56
S3S3S3			12		1*	13
S3S3S3S3				9	1	10
S3S3S3S3S3					3**	3
Total	279	59	20	11	5	374

* [S4,S3,S3,A1,S3,S3,S3]
** includes [S2,S3,S3,S3,S3,S3]

Table 12: Comparison of longest sequence and total number of occurrences of S5

Longest S5 sequence	Total number of S5's in the test history				Total
	1	2	3	4	
S5	257	8			265
S5S5		12			12
S5S5S5			0		0
S5S5S55				1	1
Total	257	20	0	1	278

Table 13: Comparison of longest sequence and total number of occurrences of U

Longest U sequence	Total number of U's in the test history					Total
	1	2	3	4	5	
U	623	62	1			686
UU		129	15	3		147
UUU			26	3		29
UUUU				7		7
UUUUU					2	2
Total	623	191	42	13	2	871

Table 14: Comparison of longest sequence and total number of occurrences of U1

Longest U1 sequence	Total number of U1's in the test history					Total
	1	2	3	4	5	
U1	601	53	2			656
U1U1		112	15	3		130
U1U1U1			23	2		25
U1U1U1U1				6		6
U1U1U1U1U1					2	2
Total	601	165	40	11	2	819

Table 15: Comparison of longest sequence and total number of occurrences of U2

Longest U2 sequence	Total number of U2's in the test history			Total
	1	2	3	
U2	64	1		65
U2U2		8		8
U2U2U2			2	2
Total	64	9	2	75

Table 16: Number of students with A sequence of length n and SFP, by school group

School group	Length of longest A sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	204	107	67	28	1			407	57%
SGB	247	124	109	28	1			509	75%
SGC	112	68	66	77	57	11	3	394	72%
SGD	72	63	48	52	57	16	2	310	78%
SGE	41	36	32	41	72	21		243	94%
SGF	194	92	125	92	10			513	84%
Overall	870	490	447	318	198	48	5	2376	74%

Table 17: Number of students with A1 sequence of length n and SFP, by school group

School group	Length of longest A1 sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	177	92	60	21				350	49%
SGB	218	119	89	25	1			452	67%
SGC	111	58	60	73	44	10	1	357	65%
SGD	74	66	46	48	44	13		291	73%
SGE	37	31	36	40	66	18		228	88%
SGF	199	80	107	65	8			459	75%
Total	816	446	398	272	163	41	1	2137	67%

Table 18: Number of students with A2 sequence of length n and SFP, by school group

School group	Length of longest A2 sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	57	3						60	8%
SGB	55	2						57	8%
SGC	51	8	2					61	11%
SGD	45	3	3					51	13%
SGE	32	2						34	13%
SGF	71	9	5	2				87	14%
Total	311	27	10	2				350	11%

Table 19: Number of students with A3 sequence of length n and SFP, by school group

School group	Length of longest A3 sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	44	1						45	6%
SGB	42	3	1					46	7%
SGC	50	6	1	1				58	11%
SGD	32	3	1					36	9%
SGE	17							17	7%
SGF	65	1						66	11%
Total	250	14	3	1				268	8%

Table 20: Number of students with L sequence of length n and SFP, by school group

School	Length of longest L sequence							Total	SFP
group	1	2	3	4	5	6	7	students	
SGA	188	60	21	15	6			290	41%
SGB	93	19	5	4				121	18%
SGC	150	75	35	17	12	1		290	53%
SGD	115	31	9	4	2			161	41%
SGE	28	4	3					35	14%
SGF	72	1		2				75	12%
Total	646	190	73	42	20	1		972	30%

Table 21: Number of students with L1 sequence of length n and SFP, by school group

School	Length of longest L1 sequence							Total	SFP
group	1	2	3	4	5	6	7	students	
SGA	153	47	17	8	4			229	32%
SGB	64	11	4	1				80	12%
SGC	131	53	22	7	5			218	40%
SGD	96	18	5	1				120	30%
SGE	12	4						16	6%
SGF	29	0	0	2				31	5%
Total	485	133	48	19	9			694	22%

Table 22: Number of students with L2 sequence of length n and SFP, by school group

School	Length of longest L2 sequence							Total	SFP
group	1	2	3	4	5	6	7	students	
SGA	55	5						60	8%
SGB	30	3						33	5%
SGC	72	8	2					82	15%
SGD	40	5						45	11%
SGE	15							15	6%
SGF	25	1						26	4%
Total	237	22	2					261	8%

Table 23: Number of students with L4 sequence of length n and SFP, by school group

School	Length of longest L4 sequence							Total	SFP
group	1	2	3	4	5	6	7	students	
SGA	50	1						51	7%
SGB	25							25	4%
SGC	72	7						79	14%
SGD	30	2						32	8%
SGE	12							12	5%
SGF	19							19	3%
Total	208	10						218	7%

Table 24: Number of students with S sequence of length n and SFP, by school group

School group	Length of longest S sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	122	16	9	2				149	21%
SGB	123	29	10	2				164	24%
SGC	94	35	8	7	1			145	26%
SGD	81	23	9	6	2			121	30%
SGE	29	10	4	2	0	1		46	18%
SGF	106	19	7	4				136	22%
Total	555	132	47	23	3	1		761	24%

Table 25: Number of students with S1 sequence of length n and SFP, by school group

School group	Length of longest S1 sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	28	2	2					32	5%
SGB	61	3						64	9%
SGC	45	10	1					56	10%
SGD	47	6						53	13%
SGE	16	3						19	7%
SGF	39	2						41	7%
Total	236	26	3					265	8%

Table 26: Number of students with S3 sequence of length n and SFP, by school group

School group	Length of longest S3 sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	73	7	1					81	11%
SGB	57	10	3	1				71	10%
SGC	51	11	4	4				70	13%
SGD	39	14	0	4	2			59	15%
SGE	10	7	2	0	1			20	8%
SGF	62	7	3	1				73	12%
Total	292	56	13	10	3			374	12%

Table 27: Number of students with S5 sequence of length n and SFP, by school group

School group	Length of longest S5 sequence 1	2	3	4	5	6	7	Total students	SFP
SGA	52	5						57	8%
SGB	65	1						66	10%
SGC	46	4						50	9%
SGD	36							36	9%
SGE	19							19	7%
SGF	47	2		1				50	8%
Total	265	12		1				278	9%

Table 28: Number of students with U sequence of length n and SFP, by school group

School group	Length of longest U sequence							Total students	SFP
	1	2	3	4	5	6	7		
SGA	169	30	3	1				203	29%
SGB	111	20	1					132	19%
SGC	134	44	5	4	2			189	34%
SGD	102	19	8	2				131	33%
SGE	43	13	2					58	22%
SGF	127	21	10					158	26%
Total	686	147	29	7	2			871	27%

Table 29: Number of students with U1 sequence of length n and SFP, by school group

School group	Length of longest U1 sequence							Total students	SFP
	1	2	3	4	5	6	7		
SGA	158	28	2	1				189	27%
SGB	107	13	1					121	18%
SGC	130	40	5	4	2			181	33%
SGD	103	20	6	1				130	33%
SGE	39	12	2					53	21%
SGF	119	17	9					145	24%
Total	656	130	25	6	2			819	26%

Table 30: Number of students with U2 sequence of length n and SFP, by school group

School group	Length of longest U2 sequence							Total students	SFP
	1	2	3	4	5	6	7		
SGA	14	1	1					16	2%
SGB	14	2						16	2%
SGC	12	1						13	2%
SGD	4	1						5	1%
SGE	4	1						5	2%
SGF	17	2	1					20	3%
Total	65	8	2					75	2%

Table 31: Sample size for calculations in Figure 6.1

Code	Gr4/5	Gr6	Gr7	Gr8	Gr9/10
A2	24	33	88	79	56
L1	469	164	153	54	0
S3	70	59	107	116	33
S5	41	34	65	44	33

Table 32: Sample size for calculations in Figure 6.2

Code	Gr4/5	Gr6	Gr7	Gr8	Gr9/10
S1	50	57	80	35	23
S3	70	59	107	116	33
S5	41	34	65	44	33

Table 33: Sample size for calculations in Figure 6.3

Grades	SGA	SGB	SGC	SGD	SGE	SGF
Grades 4-6	28	56	108	97	22	0
Grades 7-8	67	84	82	73	43	98
Grades 9-10	29	14	0	0	0	41

Table 34: Sample size for calculations in Figure 6.5

Code	8 months (T_iT_{i+1})	16 months (T_iT_{i+2})	24 months (T_iT_{i+3})	32 months (T_iT_{i+4})
L	1257	959	652	325
S	847	576	332	121
U	808	523	261	97
L1	853	667	458	240
S3	385	266	146	43
U1	757	504	260	97
A2	280	166	89	39

Appendix 9: Persistence and proximity to expertise by grade

Table 1: Proximity (%) to A1 for A2, S1 and S3 codes by grade

Grade of non-A1 test	From A2		From S1		From S3	
Grade 4/5	70	(n=23)	43	(n=42)	40	(n=53)
Grade 6	73	(n=30)	59	(n=49)	40	(n=43)
Grade 7	65	(n=74)	24	(n=68)	43	(n=70)
Grade 8	72	(n=60)	31	(n=32)	22	(n=69)
Grade 9/10	45	(n=40)	36	(n=22)	13	(n=23)
Overall	65	(n=227)	38	(n=213)	33	(n=258)
Primary	*72*	*(n=53)*	*52*	*(n=91)*	*40*	*(n=96)*
Secondary	*63*	*(n=174)*	*28*	*(n=122)*	*30*	*(n=162)*
Secondary-Primary	*-9*		*-24*		*-10*	

Table 2: Persistence-Proximity for A2 tests by grade

Grade	Persistence		Proximity	
Grade 4/5	4%	(n=24)	70%	(n=23)
Grade 6	9%	(n=33)	73%	(n=30)
Grade 7	16%	(n=88)	65%	(n=74)
Grade 8	24%	(n=79)	72%	(n=60)
Grade 9/10	29%	(n=56)	45%	(n=40)

Table 3: Persistence-Proximity for A3 tests by grade

Grade	Persistence		Proximity	
Grade 4/5	10%	(n=21)	37%	(n=19)
Grade 6	22%	(n=23)	61%	(n=18)
Grade 7	14%	(n=57)	63%	(n=49)
Grade 8	14%	(n=49)	57%	(n=42)
Grade 9/10	3%	(n=37)	56%	(n=36)

Table 4: Persistence-Proximity for L1 tests by grade

Grade	Persistence		Proximity	
Grade 4	36%	(n=179)	18%	(n=114)
Grade 5	46%	(n=290)	16%	(n=158)
Grade 6	32%	(n=164)	33%	(n=112)
Grade 7	31%	(n=153)	32%	(n=106)
Grade 8-10	39%	(n=67)	27%	(n=41)

Table 5: Persistence-Proximity for L2 tests by grade

Grade	Persistence		Proximity	
Grade 4	7%	(n=14)	23%	(n=13)
Grade 5	18%	(n=51)	21%	(n=42)
Grade 6	11%	(n=57)	51%	(n=51)
Grade 7	8%	(n=48)	32%	(n=44)
Grade 8-10	7%	(n=57)	40%	(n=53)

Table 6: Persistence-Proximity for L4 tests by grade

Grade	Persistence		Proximity	
Grade 4	12%	(n=17)	13%	(n=15)
Grade 5	6%	(n=52)	29%	(n=49)
Grade 6	5%	(n=43)	32%	(n=41)
Grade 7	5%	(n=42)	38%	(n=40)
Grade 8-10	13%	(n=23)	30%	(n=20)

Table 7: Persistence-Proximity for S1 tests by grade

Grade	Persistence		Proximity	
Grade 4/5	16%	(n=50)	43%	(n=42)
Grade 6	14%	(n=57)	59%	(n=49)
Grade 7	15%	(n=80)	24%	(n=68)
Grade 8	9%	(n=35)	31%	(n=32)
Grade 9/10	4%	(n=23)	36%	(n=22)

Table 8: Persistence-Proximity for S3 tests by grade

Grade	Persistence		Proximity	
Grade 4/5	24%	(n=70)	40%	(n=53)
Grade 6	27%	(n=59)	40%	(n=43)
Grade 7	35%	(n=107)	43%	(n=70)
Grade 8	41%	(n=116)	22%	(n=69)
Grade 9/10	30%	(n=33)	13%	(n=23)

Table 9: Persistence-Proximity for S5 tests by grade

Grade	Persistence		Proximity	
Grade 4/5	2%	(n=41)	20%	(n=40)
Grade 6	6%	(n=34)	41%	(n=32)
Grade 7	2%	(n=65)	31%	(n=64)
Grade 8	14%	(n=44)	26%	(n=38)
Grade 9/10	15%	(n=33)	43%	(n=28)

Table 10: Persistence-Proximity for U1 tests by grade

Grade	Persistence		Proximity	
Grade 4	6%	(n=33)	10%	(n=31)
Grade 5	29%	(n=141)	55%	(n=100)
Grade 6	30%	(n=147)	57%	(n=103)
Grade 7	28%	(n=188)	50%	(n=135)
Grade 8	29%	(n=140)	54%	(n=99)
Grade 9/10	26%	(n=108)	53%	(n=80)

Table 11: Persistence-Proximity for U2 tests

	Persistence		Proximity	
Overall	24%	(n=51)	56%	(n=39)

Appendix 9

Appendix 10: Regression from expertise

Table 1: Numbers of regression test pairs by the semesters of the two tests

Semester of unstable A1	Semester of regression test												Total
	Gr5-Sem1	Gr5-Sem2	Gr6-Sem1	Gr6-Sem2	Gr7-Sem1	Gr7-Sem2	Gr8-Sem1	Gr8-Sem2	Gr9-Sem1	Gr9-Sem2	Gr10-Sem1	Gr10-Sem2	
Gr4–Sem2	1	2	1										4
Gr5–Sem1		10	3										13
Gr5–Sem2			11	2	0	1							14
Gr6–Sem1				23	1								24
Gr6–Sem2					17	0	1						18
Gr7–Sem1						28	7	3	1	1			40
Gr7–Sem2							41	23	1	4			69
Gr8–Sem1								36	7	2			45
Gr8–Sem2									41	17	2	1	61
Gr9–Sem1										10	1		11
Gr9–Sem2											29	13	42
Gr10–Sem1												8	8
Total	1	12	15	25	18	29	49	62	50	34	32	22	349

Shading indicates tests in consecutive semesters 255/349=73%

Table 2: Regression rate by grade of regression test

Grade of regression test	Number of test pairs			Regression rate (%)
	Total (A1,X)	Retest (A1,A1)	Regression (A1,non-A1)	
Grade 5	46	33	13	28
Grade 6	350	310	40	11
Grade 7	641	594	47	7
Grade 8	989	878	111	11
Grade 9	775	691	84	11
Grade 10	478	424	54	11
Total	3279	2930	349	11

Table 3: Numbers of regressions by grade of unstable A1 and by school group

School Group	Grade of unstable A1							Total
	4	5	6	7	8	9	10	
SGA	0	1	1	24	6	18	2	52
SGB	0	4	4	16	7	6	0	37
SGC	3	5	8	28	24	0	0	68
SGD	1	11	21	22	17	0	0	72
SGE	0	6	8	14	2	3	0	33
SGF	0	0	0	5	50	26	6	87
Total	4	27	42	109	106	53	8	349

Table 4: Regression test codes by grade of regression test

Grade of regression test	Regression test codes										Total
	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2	
Grade 5	3	0	0	2	0	0	1	0	7	0	13
Grade 6	7	4	2	6	0	4	3	0	13	1	40
Grade 7	10	6	1	3	0	5	0	2	20	0	47
Grade 8	34	13	3	3	0	4	6	12	27	9	111
Grade 9	19	14	0	4	1	4	3	4	27	8	84
Grade 10	14	10	1	1	0	1	3	1	18	5	54
Total	87	47	7	19	1	18	16	19	112	23	349

Table 5: Regression test codes by school group

School Group	Regression test codes										Total
	A2	A3	L1	L2	L4	S1	S3	S5	U1	U2	
SGA	16	4	1	3	0	1	3	4	16	4	52
SGB	8	8	0	0	1	3	5	2	6	4	37
SGC	16	6	2	4	0	5	4	4	24	3	68
SGD	20	5	0	5	0	6	1	4	30	1	72
SGE	6	5	2	4	0	0	1	1	11	3	33
SGF	21	19	2	3	0	3	2	4	25	8	87
Total	87	47	7	19	1	18	16	19	112	23	349

Table 6: Grouped regression test codes by school group

School Group	Grouped regression test codes				Total
	A2/A3	Any L	Any S	Any U	
SGA	20	4	8	20	52
SGB	16	1	10	10	37
SGC	22	6	13	27	68
SGD	25	5	11	31	72
SGE	11	6	2	14	33
SGF	40	5	9	33	87
Total	134	27	53	135	349

Table 7: Grouped regression test codes by test before A1

Test before A1	Grouped regression test codes				Total
	A2/A3	Any L	Any S	Any U	
A1	37	3	5	28	73
A2/3	16	1	1	6	24
L	10	9	5	9	33
S	11	1	10	11	33
U	13	2	7	18	40
Total	87	16	28	72	203

Appendix 11: The project effect

Table 1 indicates the level of A1 for the two groups of tests, as well as the number of tests involved. The project effect is the difference between the levels of expertise of the two groups. The project effect is negative for both semesters in Grade 5, and positive for the Grades 6 to 10. The last column is the chi-squared statistic. Nine of the eleven semesters have a statistically significant project effect (italicised entries). To determine the approximate size of the project effect, calculations based on sample sizes of more than 50 were considered. Averaging the project effect for these 10 semesters provides a value of 11%.

Table: Calculation of a Project Effect

Semester	A1% First test		A1% Subsequent tests		Project Effect	χ^2
Gr4-Sem1	0	(n=39)	-	(n=0)	-	
Gr4-Sem2	4	(n=297)	-	(n=0)	-	
Gr5-Sem1	18	(n=269)	7	(n=145)	-11	*0.0022*
Gr5-Sem2	32	(n=210)	20	(n=337)	-12	*0.0020*
Gr6-Sem1	32	(n=149)	39	(n=520)	7	0.1097
Gr6-Sem2	30	(n=115)	54	(n=665)	25	*0.0000*
Gr7-Sem1	47	(n=416)	67	(n=495)	20	*0.0000*
Gr7-Sem2	48	(n=731)	63	(n=655)	15	*0.0000*
Gr8-Sem1	29	(n=69)	68	(n=848)	39	*0.0000*
Gr8-Sem2	46	(n=316)	64	(n=873)	18	*0.0000*
Gr9-Sem1	27	(n=22)	69	(n=839)	42	*0.0000*
Gr9-Sem2	57	(n=288)	62	(n=498)	5	0.1590
Gr10-Sem1	47	(n=38)	70	(n=442)	23	*0.0038*
Gr10-Sem2	62	(n=245)	67	(n=341)	5	0.2381
Average project effect (based on samples of 50 or more tests)					11	