**University of Cape Town**
**Department of Physics**

# First year students' understanding of measurement in physics laboratory work

**Trevor Stanton Volkwyn**

A dissertation submitted to the Faculty of Science at the University of Cape Town
in fulfilment of the requirements for the degree of Master of Science in Physics

**July 2005**

I declare that, except where acknowledged, the work contained in this thesis is my own original work, carried out with guidance and advice from my supervisors.

***Trevor S Volkwyn***
15 July 2005

# Abstract

Recent collaborative work by the physics education research group at the University of Cape Town (South Africa) and the science education research group at the University of York (United Kingdom) has produced a suite of research instruments which may be used to probe the procedural understanding of first-year physics students. The work has led to the development of a model for classifying students' reasoning about measurement in terms of theoretical constructs which have been termed the point and set paradigms. The model accounts for the ways in which students make decisions in the areas of data collection, data processing and data comparison during experimental work. A set of questionnaires was modified and used in this study to investigate mainstream physics students' understanding of measurement both before and after completing a full year physics laboratory curriculum. It was found that although the mainstream students both entered and exited their course with high levels of proficiency in applying the more formalistic rules of data analysis, very few shifted in their fundamental understanding of the concepts that underlie experimentation. The results further suggest that the laboratory course may have indeed impeded these students from developing a deep understanding of the nature of measurement and uncertainty.

# Acknowledgements

This dissertation was completed under the supervision of Assoc Profs S Allie and A Buffler.

I developed a love for teaching and physics education research under the guidance and mentoring of Saalih Allie. The many stimulating discussions, that yielded interesting ideas and points of reflection, formed the basis for much of this dissertation. I thank him for believing in me. Andy Buffler's structured supervision, expert editing and proofreading, and his clear vision about the essence of the work, was instrumental in bringing the writing process to completion. With constant encouragement, he stimulated progress throughout. I thank him for exposing me to the discipline and dedication required for academic work.

This work benefited greatly from the expert knowledge of education research of Fred Lubben, of the Science Education Department, University of York (UK). His many comments and contributions were invaluable in producing a coherent document in the end.

Dr George Malek of the Ecumenical Pastoral Institute in Cape Town is credited with salvaging my post-graduate studies. Our many hours of discussion and deep spiritual reflection provided the basis for restoring my psyche and spirit to that required for engaging life effectively.

The timely interventions of Prof. Craig Comrie (Head of Physics Department, UCT) were instrumental in ensuring the completion of this work. I thank the Physics Department for hosting me and providing all the necessary administrative and material assistance and a collegiate environment. I wish to thank especially the Physics Education Research Group for the many stimulating and thought provoking discussions that have challenged my understanding of physics and teaching practice and in particular Rudolph Nchodu, for his friendship, advice and encouragement.

Many others are acknowledged for their contribution to this work – Kerwin Ontong for technical assistance; Fiona Gibbons and Nomphelo Lungisa who assisted with administration of the project; and George Swartz and Juliet Davids for the printing and binding of the probe sheets.

I gratefully acknowledge the financial assistance provided by the National Research Foundation.

Finally, I express my deep appreciation to my parents and family who have been consistent and unselfish in their support of my studies and for tolerating my 'anti-social behaviour' for so long.

# Contents

# List of figures

# List of tables

# 1

# Introduction

The work reported in this dissertation forms part of a wider research programme undertaken at the University of Cape Town (UCT), South Africa, to investigate and interpret undergraduate students' understandings of measurement. The science education research groups in the Department of Physics at UCT and the Department of Educational Studies at the University of York (UOY), United Kingdom, have been engaged in a collaborative effort (since 1995) to develop a theoretical basis for the construction and implementation of a new introductory physics laboratory curriculum. The new laboratory course would not only facilitate the development of students' abilities in performing experimental procedures and using the tools of data analysis, but also deepen their understanding of the nature of measurement and uncertainty.

Over the past few decades there has been an increasing body of literature that focuses on the learning and teaching of physics. For example, a selection of papers that are directed primarily at undergraduate university level are listed in the review article by McDermott and Redish (1999) published in the *American Journal of Physics*. The resource article lists a large number of empirical studies under the following headings (the number in brackets gives the number of articles listed): A. Conceptual understanding in Mechanics (56), Electricity and magnetism (20), Light and optics (15), Properties of matter, Fluid mechanics and thermal physics (14), Waves and sound (6), Topics in modern physics (4); B. Problem-solving performance (12); D. Ability to apply mathematics in physics (4); E. Attitudes and beliefs of students (11); F. Reflections on research in student reasoning (4). Under Section C (Effectiveness of laboratory instruction and lecture demonstrations), only 6 papers: Gerson and Primrose (1977), Reif and St. John (1979), Long *et al.* (1986), Séré *et al.* (1993), Roth *et al.* (1997), and Allie *et al.* (1998), are listed, showing the relatively few studies completed which deal with laboratory work. At school level a number of

studies dealing with laboratory-associated issues have been reported, particularly in the United Kingdom: e.g. Millar *et al.* (1994), Gott and Duggan (1995), Millar *et al.* (1996), and Lubben and Millar (1996). A survey of the literature suggests that comparatively little work has been done in the area of student understanding of measurement, especially at undergraduate university level.

Most undergraduate level physics courses consist of a theoretical component, administered through lectures and tutorials, and a laboratory-based experimental component. It is evident that most science educators consider experimentation to be an integral part of any course that seeks to provide students with an introduction to the scientific discipline studied. The purposes of laboratory work, in the context of students' progress in learning physics, need to be clear when designing a course on measurement. Unfortunately, the literature in the fields of science and physics education research suggests confusing and sometimes contradictory reasons for including practical work in science courses. White (1996) highlights the wide spectrum of opinions and convictions held by science educators about the place of laboratory work in science courses. Importantly, he points out that no consensus exists on what the aims of laboratories should be (e.g. learning of skills, concepts, developing reasoning, deepening of understanding of the nature of science, etc.) and that there is little evidence to suggest that they are indeed effective in accomplishing these goals. Similar observations have been made by other science education researchers with many suggesting a variety of intervention strategies and approaches to enhance the effectiveness of laboratory work (see for example Gerson and Primrose, 1977; Reif and St. John, 1979; Long *et al.*, 1986; and Arons, 1993). These educationists make some valuable contributions, such as providing ideas for open-ended investigations and advising that Socratic dialogue should form the basis of the teacher-student interaction when conducting experimental work. However, a clear indication of what a practical course should aim for can only be obtained by first considering the different aspects of scientific knowledge and how laboratory work may contribute to this knowledge.

Millar *et al.* (1994), in their report on an investigation regarding the introduction of a new school science curriculum in the UK (details in next section), highlight four areas of understanding needed in the various performance stages ('actions') of scientific investigation. These stages are (i) the interpretation of what the given task involves; (ii) the completion of a set of observations or measurements; (iii) interpretation of results to draw conclusions; and (iv) the evaluation of the conclusions in light of the aims and purposes of the investigation and how these compare or contribute to the established body of knowledge. The first area of understanding, as characterised by Millar *et al.* (1994), concerns a student's *declarative* ('knowing what') knowledge (see Black, 1993) of the various science concepts involved in an investigation, which is applicable to all the

performance stages. The other three areas of understanding may be viewed as elements of *procedural* ('knowing how') knowledge of carrying out an investigation. At this point it must be stated that there is no agreed terminology for discussing procedural knowledge in the literature. A distinction needs to be drawn between 'procedural knowledge' as defined in experimental work, and the ability of students to apply algorithmic procedures when solving written problems (Larkin and Reif, 1979; Chi *et al.*, 1981). Nevertheless, the following categories are useful in characterising students' understanding. They are: (a) the 'frame' or understanding of the aims and purposes of the investigation, generally impacting on performance stage (i); (b) the ability to carry out applicable manipulative skills, including the use of instruments and performing standard procedures, necessary for stage (ii) of an investigation; and (c) the understanding of evidence which involves the different ideas students may have about the criteria used to judge the quality and validity of empirical data ('concepts of evidence' - Gott and Duggan, 1996); critically influencing all the stages of, or actions taken in, a typical investigation. It must be noted that relatively few studies have been reported focusing on students' procedural knowledge (Roth and Roychondhury, 1993; Germann and Aram, 1996) compared to the large number addressing declarative knowledge (summarised by Pfundt and Duit, 1994).

Most undergraduate physics laboratory courses consist of highly structured experiments (of the "verification type") that are intended to increase students' understanding of the concepts, laws and models (theory component – declarative knowledge) introduced in lectures (Meester and Maskill, 1995; Laws, 1996; Tiberghien *et al.*, 2001). However, studies have been reported that express serious doubt about the effectiveness of traditional experimental activities for illustrating theory or phenomena (Roth *et al.*, 1997; Kirchner and Huisman, 1998). One of the key outcomes of scientific enquiry is the establishment of theories; hence exposing students to artificial experiments in which previously established laws and theories are verified is likely to lead to serious distortions (e.g. the idea that the "perfect" experiment yields the "perfect" result or desired outcome) in students' understanding of the scientific approach to enquiry (an element of learning about science – Hodson, 1998). Etkina *et al.* (2002) have attempted to clarify and organise the aims of experiments in terms of their specific purposes – termed a 'process approach'. They suggested that laboratory exercises be structured around one of three types of experiments; those that are for illustration of new phenomena for which students have to formulate explanations (*observational experiments*), those where a prediction is verified based on a previously developed explanation of the same phenomenon (*testing experiments*), and exercises where students need to use the explanation for one phenomenon to predict another (*application experiments*). A laboratory curriculum based on the 'process approach' then also emphasises concept and model development through laboratory experiences.

In contrast to courses using laboratory work to develop concepts and models, Osborne (1996) argues that the purposes of hands-on practical work should be more strongly focused on developing the scientific approach to enquiry. Leach (1999) studied school students' abilities to grasp the interplay between theory and evidence in science. His purpose was to investigate claims (e.g. Kuhn *et al.*, 1998) that students, at early stages of educational development, are not capable of co-ordinating theory and evidence, i.e. they lacked the ability to test and evaluate knowledge claims in the light of observed data or experimental evidence. His findings suggested that students at these early stages are indeed capable of holding theory and evidence separately, and that what the students in fact lacked was knowledge of the rules for theory evaluation. In his paper, Leach (1999) expressed the conviction that engagement in the 'difficult' processes of measuring quantities and drawing conclusions in scientific investigations may lead to a deeper appreciation for the nature of theory, data and explanation in science itself.

Of significance is that the above characterisation of students' procedural understandings constitutes a distinct domain of knowledge to be learned, rather than a collection of skills to be practiced. A narrow focus of experimental work on the science concepts applicable to a field of study is inadequate if students are to improve their performance in carrying out and reporting on investigative tasks. Notably, Gott *et al.* (1999) reported on how critical procedural understanding is in defining the technical skills and abilities required by industry. The fact that both employers and employees found it equally difficult to define what constitutes this knowledge base, attests to the neglect of this critical field of science education. The consequence of the philosophy that procedural understandings form a knowledge base is that laboratory exercises need to be structured and designed to facilitate this. In light of the available research, as selectively presented above, the UCT-York research group believes that the undergraduate laboratory should be utilised to explicitly teach the fundamentals of measurement and uncertainty and through a structured intervention programme improve students' scientific approach to enquiry. Although useful and valuable, the published research resources mentioned above do not suggest a theoretical framework of student understanding that may be used as a basis for developing a research-based laboratory course. It would therefore be desirable to gain insights into the various shortcomings students may have in dealing with measurement related activities and establish a knowledge state of students' understanding of measurement at the undergraduate level.

## 1.1 Research into students' understanding of measurement and uncertainty

Séré *et al.* (1993) have reported a study on French first-year university students. The researchers wanted to gain insights into the students' conceptions about and difficulties with measurement, after receiving a theoretical course on data analysis. The teaching content of this course was aimed at student learning of concepts, e.g. precision, accuracy and uncertainty, and mathematical tools, e.g. mean value, standard deviation and confidence interval. After the course, it was observed that most students lacked a thorough understanding of the statistical procedures required in actual practical tasks. Students viewed first measurements as pre-eminent, subsequent ones used only to 'judge' the preceding ones; confidence intervals were applied to repeated measurements as if they were single observations, not viewing all repeats as an ensemble to be modelled; the concepts of precision and accuracy were not clearly distinguished, and most students took no initiative to comment on or compare results. The researchers reflected that 'calculation routines' should be avoided in favour of exercises or tasks that promote students' underlying understanding of data analysis.

In their study of first year chemistry students in the United Kingdom, Garrett *et al.* (2000) postulated that their students' misconceptions about specific aspects of data analysis ('best fit' straight line fitting procedures and the meaning of confidence limits) are related to a wider range of misconceptions about the nature and origin of "experimental error". These should be recognised and accounted for properly to effect learning in this area. They also found the same confusion with terminology as reported by Séré *et al.* (1993).

Coelho and Séré (1998) interviewed French secondary school students aged 14-17 years, to elicit their conceptions and difficulties during a measurement activity. Instead of labelling students' ideas as misconceptions, the researchers opted for categorising different conceptions as precursors (advantages) or obstacles (disadvantages) to a sound understanding of measurement, depending on the teaching and learning activities presented to students. They thus viewed their students' pursuit of a 'true value' and dissatisfaction with measurement variability, on the one hand, as being a productive precursor for attaining certain desired experimental outcomes (e.g. the need to repeat measurements and minimise variability), and on the other hand, being responsible for the belief that uncertainty may be eliminated completely if 'perfect' techniques or ideal equipment are used by 'expert' scientists. They ascribe this to a 'spontaneous deep realism'. With this approach, the view that measurement variability arises due to outside sources is seen as a precursor to an analytical

method of linking these sources to a numerical value of the uncertainty. Taking into account all the measurements is seen as the first step towards a statistical treatment of a set of varying data. One problem with practical courses is the closed nature of many laboratory tasks as recognised by Fairbrother and Hackling (1997), who claim that this stems from the epistemological view of science as a body of facts to be catalogued. Students believe in the existence of a 'right answer' to any experimental observation. When students observe variation in repeated measurements or obtain an answer different than the expected result, they believe that they have made an 'error' or mistake. These observations suggest that traditional laboratory based courses, due to the type and purposes of activities, are not designed to treat students' preconceptions as starting points for the development of ideas and concepts. Practical curricula may rather impede effective learning of measurement by entrenching problematic epistemological beliefs. Supporting these findings, it has been reported that students reason differently in a scientific context than when operating in the everyday experiential domain (Reif and Larkin, 1991). Since students view science as a collection of facts to be catalogued, they are likely to resist the uncomfortable process of establishing in their minds new concepts of a relatively unfamiliar domain, such as physics, through the same mental processes they would use when making sense of everyday phenomena. Compounding these difficulties, Reif and Larkin correctly point out that the science taught at schools differs from real science and everyday life, which leads to students developing distorted views of scientific goals and problematic ways of thinking in the scientific domain, which includes scientific evidence.

Hammer (1994) identified three categories of students' understanding of the nature of knowledge and learning. The undergraduate physics students involved in his investigation held beliefs about, (i) the structure of physics (students either thought that physics consists of a loose collection of isolated pieces of information or forms a coherent framework to be tied together); (ii) the content of physics (seen as a complete set of facts and formulae to be learnt or concepts to be understood and applied); and (iii) learning physics (either receiving and processing information or developing, refining and constructing own understandings). One conclusion of the study was that the beliefs held by the students, regarding the three categories of understanding described above, affected their success in learning physics. Elby (2001) recognised that traditional physics courses tend not to change students' epistemological beliefs concerning the nature and structure of scientific knowledge. He comments that even the best research-based reformed curricula that facilitate deeper conceptual understanding, fail to stimulate epistemological change. His fear is that students may revert to their 'ingrained' learning methodologies in more advanced courses, hence limiting their progress in learning physics. The strategy he expounds in his epistemologically focused course is to shift students from a view of distrust of common sense to one of refinement of everyday thinking. Hammer and Elby (2003) made the observation that "high school students have

formed robust but counter productive epistemological beliefs about science" (one of which is the idea that experimental investigations should yield a pre-verified value of a given quantity). Their findings are relevant to making laboratory measurements, as failure to verify a particular fact in the students' minds imply experimental or 'human' 'error'. The reports of Ryder and Leach (2000) and Leach *et al.* (2000), from a large study of data interpreted by nearly 800 students in upper secondary schools and universities in five European countries, seem to support the findings above. They found that the students ignored the central role of theoretical models in their interpretation of data and used multiple forms of epistemological reasoning. These findings need to be considered when designing curricula.

Séré *et al.* (2001), reporting on a diagnostic questionnaire study of about 400 French and Spanish students from high school to university, wanted to ascertain what knowledge students use to inform their actions in laboratory work and following on this, to what extent students' ways of dealing with data are informed by their epistemological positions. Reinforcing the findings in the preceding paragraph, Séré *et al.* concluded that the students' decisions in the laboratory were not based on consistent epistemological positions. Students' understanding of what entails a reliable measurement, their choice of measuring procedure, their chosen methods of processing data and their interpretation of processed data to draw conclusions, all stem from different epistemologies in different contexts.

Evangelinos *et al.* (1998) reported a study on undergraduate physics students' handling of experimental measurements, focussing on their perceptions of single readings. These students generally repeated measurements purely to validate a first measurement. When using what they considered to be a high precision instrument (one with a digital display), repetition was deemed unnecessary, believing that a single measurement could give the 'true value' of a measurand (the quantity being measured), a view that persisted even after instruction. In the minds of these students, readings from scientific instruments were seen as exact facts and precision associated with either the existence, or the lack, of many digits on the display. It was found that students' deeply held views about exactness and precision acted as barriers to their acceptance of uncertainty as an intrinsic property of scientific measurement.

Subsequently, Evangelinos *et al.* (2002) investigated the effectiveness of an intervention using the probabilistic approach to measurement (more on this in Chapter 5). The Greek first year university students involved in the study were categorised as being "exact", "approximate" or "interval" reasoners based on their views of the relationship between the measurand (variable to be measured – a theoretical construct) and a single reading (the datum). The researchers found that most

students believed that a 'good' single measurement represented an exact value. Students who realised that an ideal result is not obtainable opted for reporting a single measurement as an approximate value. Measurements were only reported as intervals when students considered them to be 'really bad'. It is clear from this that students tend to connote spread with degrees of experimental imperfections of whatever origin. The intervention, as administered by Evangelinos *et al.*, resulted in students understanding the fundamental difference between exact and uncertain quantities better, whilst learning how to apply the concepts of uncertainty and probability to single measurements.

Masnick and Morris (2002) reported a survey using interviews to establish how the characteristics of data sets influence the way students compare two sets of data. Students were more confident of their conclusions and predictions when the data sets had large sample sizes and more certain about the difference in results when the data sets had fewer overlapping data points. In addition, students based their conclusions on criteria related to comparison between data points and the means of the sets of data points. The variability or outliers of the data sets, qualities of the experimenter, or the particular apparatus used, seemed not to have influenced many students in making conclusions when comparing data.

Rather surprisingly most studies on measurement and uncertainty at university level were carried out in countries other than the USA. However, recently two noteworthy studies were completed by researchers at North American universities (Deardorrff, 2001; Lippman, 2003).

One of the most extensive studies on physics students' perceptions of measurement and uncertainty was carried out recently by Deardorff (2001). The study used both qualitative as well as quantitative procedures during the analysis. This included written surveys and interviews, as well as analysis of laboratory reports. The written and observational surveys were based on both known instruments, such as the Laboratory Procedures Questionnaire (Allie *et al.*, 1998), and new instruments that were developed and validated for specific purposes. Both "expert" and student perceptions were solicited in the study. Most of the data in this study were gathered at North Carolina State University (NCSU) with students majoring in engineering. Two samples of data were also taken at the University of North Carolina at Chapel Hill (UNC) and the University of Hokkaido in Japan.

Lippmann (2003) reported a study that evaluated an intervention, named the Scientific Community Laboratory (SCL), for teaching measurement and uncertainty to physics undergraduate students in the United States of America (USA). The approach of the SCL is to encourage students to use their

everyday reasoning skills when making decisions for the stages of data collection and data interpretation in the laboratory. The laboratory tasks are designed so that measurement 'frames' (mind-sets) are explicitly created which illustrate to students the usefulness of their everyday thinking when conducting experiments in the scientific laboratory. After following the SCL course, a large proportion of the students understood the use of intervals when comparing data sets. This result once again demonstrates the importance of establishing students' prior knowledge when enrolling for a physics course that involves practical work, and the positive outcomes that could follow if due cognisance is paid to their existing views on measurement in different contexts.

Extensive studies focussing on laboratory work at high school level were completed in the 1990's in the United Kingdom. The Procedural and Conceptual Knowledge in Science (PACKS) Project was initiated in the context of the introduction of a new national curriculum in England and Wales (Millar *et al.*, 1994) and looked into the effectiveness of school laboratory programs. One of the stated aims of the PACKS project was to develop a model linking students' performance of investigative tasks to their understanding about measurement. The actions and responses of children, 9-14 years of age, were solicited by means of given investigative tasks and diagnostic questions ('probes') to elicit aspects of the children's understanding of science concepts and procedures. It emerged that the children's performance in investigative tasks were determined not only by their understanding of the relevant science concepts, but also by their 'frames' (conceptions of the aims and purposes of investigations) and understanding of evidence (ideas and conceptions about the quality and validity of empirical data).

The findings prompted further research (a second phase of the PACKS project) into students' understanding of the fundamental ideas of validity and reliability of measurements (Lubben and Millar, 1996). The researchers undertook the survey of English secondary school and pre-university students' understanding by administering a set of written (paper and pencil) diagnostic probes, designed around a wide range of experimental settings. A model (see Table 1.1) for the progression of student ideas about measurement was suggested, with the levels ordered in terms of increasing cognitive sophistication. The authors stress that progression through the levels does not necessarily reflect students' progressive learning paths. This model does however provide a framework for classifying students' actions during measurement activities in terms of the underlying ideas about measurement.

*Table 1.1: Model of progression of ideas concerning experimental data. (Adapted from Lubben and Millar, 1996).*

| Level | Students' view of the process of measuring |
|---|---|
| A | Measure once and this is the right value. |
| B | Unless you get a value different from what you expect, a measurement is correct. |
| C | Make a few trial measurements for practice, then take the measurement you want. |
| D | Repeat measurements till you get a recurring value. This is the correct measurement. |
| E | You need to take a mean of different measurements. Slightly vary the conditions to avoid getting the same results. |
| F | Take a mean of several measurements to take care of variation due to inaccurate measuring. Quality of the result can be judged only by authority source. |
| G | Take a mean of several measurements. The spread of all the measurements indicates the quality of the result. |
| H | The consistency of the set of measurements can be judged and anomalous measurements need to be rejected before taking a mean. |

Millar *et al.* (1999) constructed a 'map' of learning outcomes in the area of a scientific approach to enquiry. Students should demonstrate proficiency in: (a) setting up a standard piece of apparatus and carrying out standard procedures; (b) planning an investigation to address a given question; (c) collecting, processing and comparing data; (d) evaluating data to support a conclusion; and (e) communicating the results of experimental work. Students' different understandings of concepts such as the validity and reliability of measurement results underlie decisions made during designing and planning experiments (learning outcome (b) above) and data manipulation (learning outcome (c) above). The UCT-UOY studies have concentrated on students' understanding of measurement and uncertainty, and hence deal with the investigative stages of data collection, data presentation and data comparison (learning outcome (c)).

Until recently, work done at UCT has been focused primarily on first year students registered for the General Entry to Programmes in Science (GEPS), formally known as the Science Foundation Programme (SFP). As a consequence of the past racially segregated schooling system in South Africa, a large majority of black students have educationally disadvantaged backgrounds. They come from generally poor socio-economic backgrounds; have had deficient schooling which includes poor facilities and inappropriately qualified teachers; had experienced instability due to political factors; and in the context of tertiary studies, English is not their first language. GEPS is

an extended, structured 4-year BSc degree programme, in contrast to the regular 3-year B.Sc. programme for mainstream or "direct entry" (DE) students, and provides educationally disadvantaged students, who do not meet the mainstream entrance requirements, an opportunity to acquire a university science degree which otherwise would be denied them.

An introductory physics course was developed for GEPS students at UCT (Allie and Buffler, 1998). The course was designed with the intention to empower students by addressing their educational needs in a physics context, drawing on research applicable to introductory physics curricula (e.g. Heller *et al.*, 1992; Hestenes, 1987; McDermott, 1991; and Van Heuvelen, 1991). Students are equipped with the practical tools, skills and procedures deemed required to overcome the demands of a physics curriculum. Laboratory work is a fundamental element of the GEPS and mainstream physics curricula at UCT. Allie *et al.* (1997) describe how laboratory reports are used to teach GEPS students to communicate the results of a scientific investigation effectively and coherently. The students' difficulties with language and concepts are addressed by utilizing physics practical exercises in improving their laboratory report writing skills. This approach, of explicitly teaching the communication of science, was supported by a study in which the laboratory reports of GEPS students were analysed to gather information on how the students carried out and communicated experimental activities (Campbell *et al.*, 2000). The results showed that the content and coherency of the students' reports were influenced by their perceptions of the purpose of the task, their understanding of the laboratory procedures involved and their knowledge of the formal way of structuring and presenting a scientific report. The suggestion was made that integrating procedural understanding and scientific communication skills in laboratory work would enhance students' understanding of measurement.

The laboratory course also introduced the formal aspects of measurement and data analysis explicitly (Allie and Buffler, 1998). The developers of the GEPS first year physics laboratory course were interested in evaluating the effectiveness of the course and identifying what the problem areas for students are in terms of their ability to engage effectively with measurement and experiments. Knowledge of the students' understanding of the nature of experimental evidence was thus crucial. The authors reported that it was evident that students experience great difficulty when encountering the concepts underlying measurement and experimentation. The course designers observed that the students became more proficient at applying the formal rules of data analysis, but that it did not necessarily indicate a deeper level of understanding of the fundamentals of measurement. To address these concerns, a survey was undertaken to explore novice university students' understanding of measurement (Allie *et al.*, 1998).

A new set of probes, based on those used in the PACKS project (Lubben and Millar, 1996), were designed and validated and used to collect data for the study as reported by Allie *et al.* (1998). A set of six probes was administered to a sample of 121 GEPS physics students during the first year of study at university. Since the methodology and design of the probes are identical to the ones used in the present work, detailed descriptions are provided in the next two chapters. In summary, students' ideas about data collection (three probes), data processing (one probe) and data comparison (two probes) were explored. These ideas that are held by students are important as they impact on the decisions made and conclusions arrived at during the stages of experimental work (see Millar *et al.*, 1994). Student responses were coded and categorised using a coding scheme that was validated by interviews. The analysis procedure involved grouping response categories according to the underlying reasoning. The relationships and consistency between the types of reasoning used in the areas of data collection, data processing and data comparison, were investigated by looking at sets of probes together for individual students. The main criteria used by students when making decisions at various stages of measurement and data analysis were identified, yielding classification schemes for the various issues being explored by the probes.

By inspecting students' written justifications for their decisions made about repeating time and distance measurements (data collection stage), a few types of reasoning emerged. Inferences were then drawn from these types of reasoning regarding students' underlying views and understanding about the nature of measurement. Where no purpose was seen in repeating measurements at all, students either did not know how to deal with variation or more likely believed that one 'good' measurement adequately represents the measurand (the quantity being measured). In responses that indicated repeating for either practising and perfecting experimental techniques, identifying a recurring value, or confirming the first or correct reading; comparisons were being made between individual readings and/or judgements made about one of a set of readings. A range of measurements taken is thus not viewed collectively as an ensemble of data points that need to be modelled by a few characterising parameters. It was deduced that for these students, repeating a measurement procedure is a search for the true value of the measurand, and that a single obtained reading could represent it. On the other hand, more than half of the cohort of students evidenced that the purpose of repeating is to obtain a mean value. However, since significantly more students wanted to obtain a mean value for time than distance measurements, there is an indication that students may have based their decision to repeat on criteria other than investigating the dispersion in obtained data and hence the modelling thereof. Only a small group of students indicated that repeats are required to observe and establish the spread in the data. Importantly, two underlying concepts emerged as the basis for students' decisions to repeat, either the measurand can be represented by a single 'true value' or a 'spread' of values.

Responses to the two data comparison probes provided insight into the consistency of students' use of 'spread reasoning'. These probes confronted students with the notion that a set of data points forms an ensemble that may be modelled by two theoretical constructs, a mean and a measure of the dispersion of the data. Students were required to make decisions on the quality (better or not) and compatibility (agree or not) of two data sets that consist of a series of readings together with their calculated means. The main finding was that even though a large proportion (about half) of students correctly concluded that a smaller dispersion in data values implies greater precision and hence a better estimate of the measurand, less than a third of these same students used the criterion of overlapping spreads to judge whether the results of two data sets are in agreement. By far the majority of students used purely subjective notions of 'closeness' of the averages to compare data ensembles. Even students who used 'spread reasoning' by indicating the calculation of a mean when answering the probe dealing with repeating of time measurements, did not conceptualise a notion of spread to be used together with the mean to characterise a measurement result. Therefore, only 15% of the total sample of students were regarded as using 'spread reasoning' consistently.

The types of reasoning that emerged with the cohort of first year university students were compared with the levels in the Lubben-Millar model of progression of understanding about measurement (Table 1.1). Only few students could be classified within levels A, C and D, and then not consistently. Although many students evidenced reasoning indicated by levels F, G, or H, those who could be classified as consistent 'spread reasoners' demonstrated greater sophistication than allowed for in the model. The researchers thus suggested that the model be extended to include an additional higher-level category (I), to identify those students who understand that a mean together with a measure of spread of a data set form an interval that may be used to judge the consistency of data ensembles. The language usage of this group of students was found to be typically haphazard, despite being classified as advanced reasoners according to the extended Lubben-Millar scheme. The responses showed that students used terms reflecting collected and computed data such as 'measurement', 'calculation', 'result', and 'value' interchangeably. Students evidenced much confusion in the use of terminology such as 'spread', 'error', range', 'uncertainty', 'precision', and 'accuracy' in their responses. The researchers do not attribute this to linguistic difficulties (students' backgrounds) alone, believing that it is more likely linked to a lack of understanding of the nature of measurement and uncertainty in the minds of the students. Evidence for this is that the vast majority of students saw the need to repeat measurements in order to limit the 'random error', and yet just over half of the sample argued at the same time for repeating to get closer to the 'real' or 'correct' value for time or distance measurements.

## 1.2 The point and set paradigm framework

The Lubben-Millar classification model described previously is a descriptive schema that ranks students' actions during the measurement phase of laboratory work according to the level of sophistication. The model however does not provide explanations for students' responses to various experimental situations nor does it theorise about the foundations of particular procedural routes chosen. Subsequent work by the UCT-York collaborative group (Allie *et al.*, 1998) found strong linkages between students' responses and their understanding of a measurement as either providing a single 'true value' or a 'spread' of values to be modelled. The constructs of point and set paradigms (see Table 1.2) were then defined and considered to account for these two 'views' of measurement. The description of the paradigms that follow below encompasses all the response types observed and is a summary of the definitions found in Lubben *et al.* (2001) and Buffler *et al.* (2001).

The **point paradigm** (Table 1.2) is characterised by the notion that each measurement could potentially produce the correct (true) value of the measurand. Variance from an expected result is caused by erroneous or uncontrollable factors. Consequently, individual measurements of a data set are viewed as independent of the others, not collectively. Adherents of the point paradigm also typically believe that a single measurement, performed expertly with good equipment under ideal conditions, adequately establishes the true value. However, when confronted with an ensemble of readings or data points with dispersion, representations of the measurand or a trend in the data are based solely on individual measurements or data points, such as selecting a recurring value in a data set, choosing only data points that fall exactly on a line when modelling a trend on a graph, and one-to-one comparisons of data values between different data sets.

The **set paradigm** on the other hand is characterised by the notion that each reading is an approximation of the measurand, knowledge of which can never be complete or perfect in principle. The measurement process is viewed as providing information about the measurand, and hence all available data is used to construct distributions from which a best estimate of the measurand and an interval of uncertainty are derived. In introductory level laboratory practicals, the best approximation of the measurand will either be the reading itself, in the case of a single reading, or the calculated average value of a set of repeated readings. Derived (combined) uncertainties are considered together with the best approximation to form confidence intervals, which are then used to make comparisons between different data sets.

*Table 1.2: The point and set paradigms.*

| **Point Paradigm** | **Set Paradigm** |
|---|---|
| The measurement process allows you to determine the true value of the measurand. | The measurement process provides incomplete information about the measurand. |
| "Errors" associated with the measurement process may be reduced to zero. | All measurements are subject to uncertainties that cannot be reduced to zero. |
| A single reading has the potential of being the true value. | All available data are used to construct distributions from which the best approximation of the measurand and an interval of uncertainty are derived. |

The fundamental difference between the two paradigms is that conclusions about the measurand are drawn directly from the individual data points when the point paradigm is used, while when the set paradigm is employed, the properties of the distribution constructed from the whole ensemble of available data informs knowledge of the measurand.

The framework of point and set paradigms have the potential of explaining students' measurement actions and reasoning; hence the UCT-York research group needed to establish the extent to which their paradigmatic model was useful for interpreting students' ideas about measurement and uncertainty. A study was undertaken, extended to a more diverse student group (GEPS and mainstream), to survey students' ideas in terms of point and set paradigms at the beginning of their academic year before instruction (Lubben *et al.*, 2001). A set of probes, identical to and extended from those used in the earlier Allie *et al.* (1998) study, was again used to explore students' understanding of measurement for the area of data collection (three probes dealing with the reasons for repeating distance and time measurements), data processing (one probe dealing with the processing of a series of repeated measurements, another the modelling of a straight line trend of a number of plotted data points), and data comparison (two probes dealing with the quality and compatibility of two data sets consisting of readings with their calculated means). The coding of the probes this time was done according to the definitions of the point and set paradigms in order to investigate the students' use of the two paradigms in the different measurement-related situations.

Individual students were identified as users of the point or set paradigms in a particular area by combining their responses to all the probes that deal with that area. Consistent use of paradigms within an area of measurement as well as across areas was determined in this way. The analysis showed that nearly two thirds of the students used the point or set paradigm in a consistent manner for decisions on measurement for data collection. There was also a good correlation between the

types of reasoning, point or set, used for data collection and that used when processing data. It was noted that the reasons for repeating measurements, the ways of dealing with a series of repeated measurements and the modelling of a straight line trend in plotted data points are all rooted in the same common construct, either the point or set paradigm. Therefore, it was reasonably concluded that the constructs of point and set paradigms form a useful framework for the interpretation of students' decision-making processes during investigative activities in the laboratory, particularly data collection and data processing. Further, only very few student responses were deemed uncodeable using the paradigmatic model for analysis of the probes, giving greater credence to the analysis method.

In contrast to the Lubben-Millar model of progression (Lubben and Millar, 1996), that links a series of measurement actions to progressive levels of understanding (Table 1.1); the point and set paradigms provide a direct classification of students' understanding of measurement. However, students' decisions were related to the procedural context of a probe, using the point and set paradigms alternately for different probes, e.g. students who were classified as consistent point reasoners (for repeating distance probes, etc.) but using the set paradigm as an exception to deal with time measurements. The use of point or set reasoning thus depended on the measurement context. In contrast, an extensive study of the Assessment Performance Unit, Department of Education and Science in the UK (APU, 1988), suggested that procedural abilities were transferable. Song and Black (1992) however reported that practical performance depended on the conceptual demand of the science context and the scientific-versus-everyday context of investigative tasks. The findings of the UCT-York group thus seems to contradict that of the APU (1998) study and further suggest that measurement decisions also depend on the measurement context of any given task. In addition, it became clear that when higher levels of measurement demand were placed on students, such as requiring decisions based on the degree of dispersion in a set of data points as an indispensable attribute of the data, set reasoning was maintained only at a low cognitive level. Students recognised the spread in the data but used only the mean to represent the data. Students, who displayed set reasoning consistently in the areas of data collection and data processing, could not continue doing so for data comparison. This is consistent with the findings of Gott and Duggan (1995) that data interpretation (of which data comparison is a part) is more demanding than experimental design, data collection and data presentation.

In summary, the point and set paradigmatic model was shown to provide a sound basis for inferring students' understanding of measurement from their actions and written justifications (reasoning). However, the use of either paradigm depended on the measurement context and consistent internalised use of set reasoning broke down for higher-level tasks. Further, some students used

both point and set reasoning in a fragmented manner, in other words their actions were not always coherent with their stated reasoning (as also noted by Germann *et al.*, 1996). A given example are those students that expressed the need to repeat readings to establish a mean but then chose a recurring value to represent a data set when an open-ended choice was provided. Some students displayed set reasoning by describing an appropriate fitting procedure to model a trend in plotted data points but then contradicted this by drawing a line segment through particular data points, an action informed by the point paradigm. It was therefore thought helpful to differentiate between reasoning about measurement and measurement action for each of the two paradigms; the result of this exercise is presented in Table 1.3.

The survey done by Lubben *et al.* (2001) showed that students could draw actions and reasoning from the point or set paradigms on an *ad hoc* basis, depending on the laboratory context. There are thus four broad categories into which students may be classified in terms of both their actions and reasoning, as illustrated in Figure 1.1 below. Students who draw their actions and reasoning purely from the point or set paradigms, as described in Table 1.3, are located in the bottom left and upper right regions respectively. A third grouping are students who use the tools and actions of the set paradigm by rote, the upper left region. They display proficiency in using the tools of statistical data analysis (see Table 1.3), but their underlying reasoning places them in the point paradigm. Lastly, in the bottom right region are those students who characteristically use actions associated with the point paradigm, but evidence reasoning that is compatible with a coherent "set" theoretical view of measurement. These students have not yet mastered the operational tools and procedures of data analysis.

With the paradigmatic framework as a basis, the UCT-York group therefore views the broad purpose of laboratory instruction, particularly for introductory level physics courses, as developing students' understanding of scientific measurement by facilitating a shift in paradigm usage. The extent to which a course is successful in effecting such a shift may be gleaned from the proportion of students who are located in the upper right region after instruction.

*Table 1.3: Actions and reasoning associated with the point and set paradigms.*

| **Point paradigm:** | | |
|---|---|---|
| **Measurement phase** | **Action** | **Reasoning** |
| Data collection | No repeating of measurements is necessary, or repeat to find recurring value, or repeat for practice. | A measurement leads to a single, "point-like" value rather an contributing to an interval. Only one good measurement is required. |
| Data processing (Calculation) | A single (best) measurement , e.g. the recurring value, is selected to represent the true value. | Each single measurement is independent of all others and can in principle be the true value. |
| Data processing (Straight line graph) | All points joined by multiple line segments or a single line drawn through selected data points. | The trend of the data is best represented by selecting particular data points which describe the desired trend. |
| Data set comparison | A value-by-value comparison of the two sets, or comparison based on the "closeness" of the means (if given). | No basis for the need to repeat measurements therefore comparisons made on the basis of the closeness of individual points. |
| **Set paradigm:** | | |
| **Measurement phase** | **Action** | **Reasoning** |
| Data collection | Repeating of measurements of the same quantity is necessary as a consequence of the inherent spread in data. | Each measurement is only an approximation to the true value and that the deviation from the true value is random. A large number of measurements are required to form a distribution that will cluster around some particular value. |
| Data processing (Calculation) | A set of measurements is represented by theoretical constructs, e.g. the mean and standard deviation. | The best information regarding the true value is obtained by combining the measurements using theoretical constructs in order to characterise the set as a whole. |
| Data processing (Straight line graph) | All the measurements taken into account by a least squares straight line fit to all the data. | The best graphical representation of series of measurements is obtained by modelling the trend of the data. |
| Data set quality | For the same number of measurements, the better measurement is chosen to be the one associated with the smallest standard deviation. | The standard deviation is related to the precision of the measurement. |
| Data set comparison | The agreement of two measurements is related to the degree of the overlap of their intervals. | The mean and standard deviation define a confidence interval that is related to both the best estimate and the reliability of the measurement. |

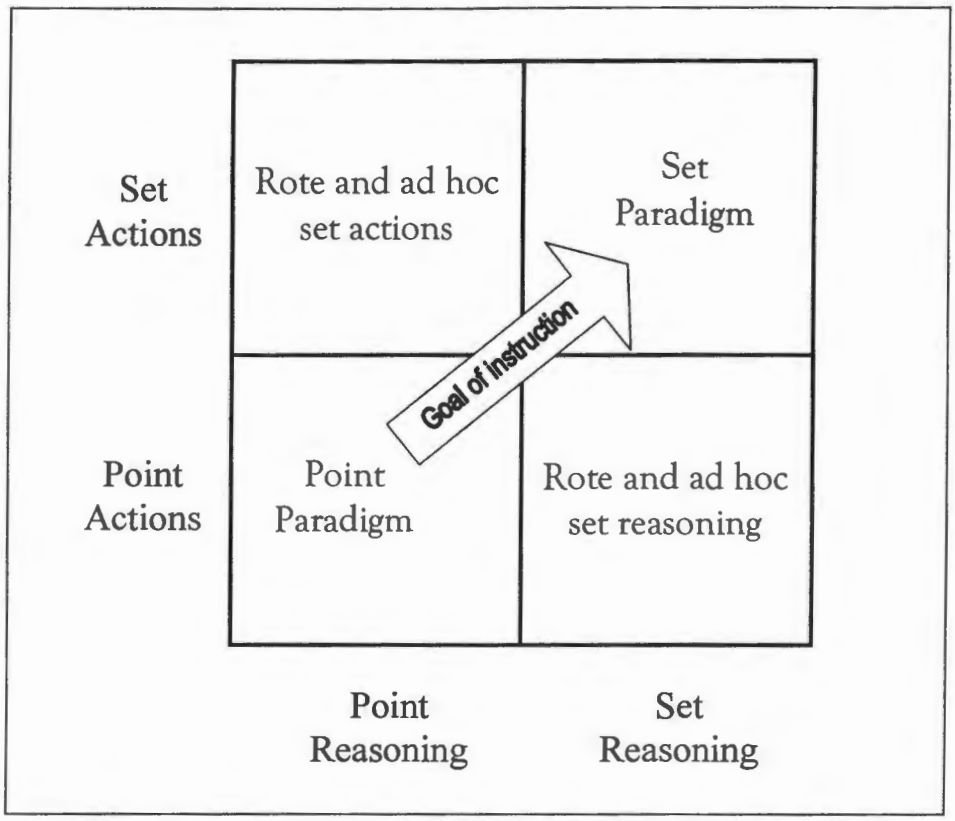



*Figure 1.1: The goal of instruction in relation to the point and set paradigms.*

## 1.3 Evaluation of the GEPS laboratory course

With the validated framework of point and set paradigms, the next phase of the research at UCT was to evaluate the GEPS physics laboratory curriculum in terms of its effectiveness in moving students toward adopting the set paradigm in the various stages of measurement in the physics first year laboratory. The work published by Buffler *et al.* (2001), which reported on aspects of this study, explored the development of the special access students' views about measurement and scientific evidence by administering the previously validated instrument (as presented in *e.g.* Allie *et al.*, 1998) both prior to and subsequent to completion of the course.

GEPS students, as explained earlier, generally use English as a second language and have had little or no exposure to "hands-on" practicals at school. The laboratory-based experimental component and the communications skills component of the GEPS course (detailed in Allie and Buffler, 1998) were thus designed and structured to address the educational needs of the students. Activities were targeted at increasing students' understanding of the measurement process, developing their skills in using various measurement instruments and guiding them to proficient application of the tools of data analysis. It was felt that fostering the students' skills such as planning and executing

structured experimental tasks and writing coherent scientific reports would not be effective with the usual "recipe-style" laboratory practicals. Consequently, typically styled laboratory manuals that contain detailed instructions for performing practical experiments were dispensed with in favour of tasks that posit authentic contextual problems that require resolution through undertaking an experimental investigation. The assessment instrument developed gave clear feedback to students by indicating explicitly the experimental and report-writing aspects that needed further attention. The approach of the course was basically to gradually build students' knowledge and understanding of measurement by the considered introduction of various concepts and deliberate focus on particular experimental skills.

The extent to which GEPS students' understanding of measurement deepened with respect to the areas of data collection, data processing and data comparison, in terms of the adoption of the set paradigm, was investigated by comparing their responses to probes before any intervention to those after the laboratory course. Significant shifts were found to have occurred in the use of paradigms for the areas of data collection and data processing. When considering individual students' responses to all the probes dealing with the reasons for repeating measurements (data collection), more than half of the cohort of students used the point paradigm consistently before the laboratory course and just over one-fifth after instruction, whereas the proportion of students who appeared to have adopted the set paradigm in a consistent manner increased from one in twelve to over two thirds. The fraction of students using the point paradigm consistently in answering the probes dealing with the use of procedures to represent series of data points (data processing), decreased from over three quarters at the start to just over one-eighth after the course and those that used the set paradigm increased from 7% to 43%. The question arose as to what extent students embraced the set paradigm. It was realised that the need to repeat measurements to calculate a mean and the fitting of a straight-line trend to a set of plotted data points may have been rote learned application of formalistic procedures and actions. Closer inspection of responses to the data collection probes revealed that the largest shift from the use of a point to set paradigm was for those students who initially indicated the need to repeat measurements to identify a recurring value (point) but later chose to calculate the mean (set). Previously, the studies have shown that students see the determination of a mean as a requirement of an experiment, and consequently, an ensemble of readings necessarily need to be generated through repeating measurements. Analysis of the data comparison probes provided the means of gauging how deeply students' understanding of the set paradigm was internalised.

The school science curriculum in South Africa does not require emphasis to be placed on the ways of dealing with spread in repeated experimental measurements, the 'average' being the only

construct introduced to represent repeated observations. It was thus not surprising that none of the GEPS students at entry used the set paradigm in a fully internalised manner. Most of the students used only the mean to compare two data sets with very few giving any recognition to the spread. After the laboratory course, the large majority of students (70%) still did not use the set paradigm consistently across all the data comparison probes. They had not internalised set reasoning in any fundamental way. When looking at the consistency of use of the point or set paradigms in all three areas of measurements (data collection, data processing and data comparison) after the intensive laboratory course, only about one-fifth (21%) of the students were found to have based their decisions consistently on an internalised set paradigm, with three quarters of the students still inconsistent in their actions and/or reasoning. An additional probe, included only in the post-instruction questionnaire, required from students a comparison between two measurement results presented in the formal manner of a mean and a standard deviation of the mean. Contrasting students' use of paradigms for this probe against that used for the previous probes showed that fewer than half of the students that applied a set action correctly appeared to be firmly rooted in the set paradigm (less than a quarter of the total sample). Three quarters of the students either did not reason according to the set paradigm or appear to have applied the correct action by rote or in an *ad hoc* way. There was thus no relationship between students' ability to apply the formalistic rules of overlapping intervals and their underlying understanding of the statistical nature of measurement. The results for the GEPS cohort demonstrated that students who have learned how to deal with data when presented in the formal way did not by implication develop the commensurate conceptual understanding of the underlying principles of set reasoning.

In summary, the GEPS laboratory curriculum seemed not to be particularly effective in shifting significant numbers of students to adoption of the set paradigm in a fundamental way when dealing with measurement in a scientific context. The findings suggested that the laboratory course, although successful in its aims of teaching students the formal procedures of data analysis, was not able to provide the necessary links between the nature of measurement and the techniques for processing data. Students successfully drew on the set paradigm for certain routine actions but resorted to the point paradigm when dealing with aspects of measurement that required deeper reasoning. In addition, it was found that students, who displayed consistent set reasoning, did not necessarily understand the operational tools of data analysis. Significant to this work, the observation was made that school learned algorithms for handling data might have hampered the development the students' understanding of measurement. Would the findings above be substantiated in a study of direct-entry students who are generally better prepared for university based on their higher achievement at school level?

## 1.4 The present work

Having found that the GEPS laboratory course (pre-2003) was largely ineffective in moving the special access students to a deeper understanding of the tools and concepts that underlie measurement activities in the laboratory, it became desirable to establish the situation for direct-entry students who intend majoring in physics at UCT. Do higher levels of preparedness translate to students more easily attaining the desired outcome of developing a deeper and more fundamental understanding about measurement? The work, presented in this dissertation, focuses on direct-entry students enrolled in the Science Faculty at UCT and investigates the understandings of these initially more advanced students both before and after their first year physics laboratory curriculum by applying the research instrument as used in the studies reported on previously (Allie *et al.*, 1998; Lubben *et al.*, 2001).

### 1.4.1 The mainstream laboratory curriculum

The laboratory course, completed by the students in the year (2000) in which the data for this survey were collected, consisted of 3-hour afternoon sessions once a week with a frequency of two sessions for every three academic weeks. The students completed a total of ten experiments for the year that were closely related to the theory introduced in lectures, mostly in the format of verifying various laws or arriving at well known values for quantities such as the acceleration due to gravity. Two projects were completed, one in each of the two 12-week semesters, that involved more investigative aspects of experimentation. A laboratory examination was written at the end of each semester, for which students were expected to have mastered the skills of designing and planning investigations, demonstrated proficiency in collecting and processing data, and evidenced skill in experimentation by arriving at numerical values for quantities to high degrees of accuracy relative to pre-determined values known by the examiners. Since the marks for these two laboratory examinations are heavily weighted in the final assessment mark for the full-year physics course, students focussed much of their attention on the generation of 'accurate' results when performing experiments.

The first two weeks of the course were dedicated to the development of various 'experimental skills'. These included the introduction of various measuring apparatus with exercises to allow students to practice and develop their skill in handling and using various instruments such as the vernier callipers, micrometer screwgauge and digital stopwatch. The two sessions started in the form of traditional lectures to introduce ideas after which the students moved to the laboratory area

to engage with the equipment. Another aspect dealt with in these introductory sessions was the basic structure of a 'laboratory write-up', after which it received no further attention for the rest of the year. It must be noted that this is not at all similar to the full laboratory report referred to earlier; the 'write-up' is basically a shortened version of the full report that focuses on the collection of data, tabulated results, plotted graphs, brief analyses and a conclusion that normally serves only to verify a stated aim. Importantly, these sessions were used to introduce students to the general approach of the course, the 'frequentist' method of dealing with scatter in data. An exercise involving radioactive decay formed the basis for introducing ideas like 'random fluctuations' and the 'normal (Gaussian) distribution' and how physicists model 'randomness'. Various 'rules of thumb' were introduced to students as ways of dealing with different measurement situations, e.g. the 'least-count' method of reporting the 'error' on a single measurement and the number of repeated readings to take for different cases (e.g. three for distance measurements with a metre rule, etc.).

All of the experiments were presented to students in the form of a 'cookbook' type laboratory manual. The language used in this manual is technical and the style terse. Appendices contained highly technical accounts of the various apparatus students would encounter during the course. The description of an experiment as it appeared in the manual contained a stated aim, a 'recipe' style method section that is essentially a list of instructions, and an analysis or theory section that instructs the reader how to process the data to a final conclusion. Students generally followed these manuals 'to the letter' with their write-ups closely mirroring the contents of the text in the account given in the manual.

Roving demonstrators, all post-graduate students in the Physics Department, were responsible for guiding students during their laboratory afternoons and to assess their write-ups, giving 'impression marks' out of twenty. It was left entirely up to the demonstrators to decide on the criteria used to arrive at a mark with selective feedback. A common complaint made by students was that different demonstrators gave different marks for similar work (students worked in pairs, keeping partners for the whole year). The markers were free to write brief comments on the workbooks, but with no clear guidelines. It is thus clear that students probably did not learn much from inspecting their returned write-ups, which were generally returned to them only at the start of their next session when an experiment would be performed. The communicative and investigative aspects of experimentation therefore seem to have been neglected in the form the mainstream laboratory course was structured and delivered to students.

### 1.4.2 Profile of the mainstream students at UCT

A total of 113 individual students completed the written questionnaires; however, 100 students participated in the pre-instruction test and just 66 in the post-study. This meant that only 53 students completed both the questionnaires, before and after the course. Although this can be explained partly by the fact that some students were absent for either or both of the tests, the main reason for the drop in numbers of students by the end of the course attests to the high attrition rate of the mainstream course. The full year course was split into two courses, one for each semester. Students who failed the mid-year examination were not allowed to continue with the mainstream major course and could opt to register for a second semester half-course in physics that ran six months out of phase with the full-year courses. This study reports only on the sample of students who completed both the pre- and post-questionnaires. This cohort of students thus provides an insight into what the best possible scenario is in terms of student ability and performance. A study of their understanding of measurement is thus a valuable exercise in investigating how traditional physics laboratory courses impact advanced students' understanding and reasoning when completing measurement related tasks.

Of the 53 students that make up the sample, 45 (85%) of the students indicated that they studied English as a first language at high school, with 38 (72%) using English as their home language. Other home languages used by this group of students were Afrikaans (3 or 6%), African languages (7 or 13%), and foreign languages including German (1), Portuguese (1), Chinese (2), Polish (1) and Gujarati (1). It is thus clear that the sample as a whole would not have found English as the language of instruction at UCT problematic. Nearly three quarters (72%) of the sample studied were male, showing that in their study year, significantly more males than females wished to take physics as one of their major subjects.

Access to the Science Faculty is determined chiefly on performance in matriculation examinations at the end of the school curriculum. Points are awarded for symbols (8 points awarded for an A symbol on the Higher Grade achieved in the matriculation examination, 7 for a B, 6 for C, 5 for a D, etc.) achieved in each of the various subjects taken by the students. Allie and Buffler (1998) provided a detailed explanation of the points system and the criteria for acceptance. The points are added and the total is used as a selection pointer. To gain access to the mainstream physics course, students would need to score good marks in both Mathematics and Physical Science as the points awarded for these two subjects are doubled in the total calculation. All the students taking part in the study studied mathematics and physical science on the higher grade. The average mark attained by these students in each of the two subjects were above 70% (equals 7 points); Mathematics (7.2)

and Physical Science or Physics (7.1). This performance level is much higher than the average for the intake of the GEPS programme which differs from the mainstream by about 10 points for all subjects counted (Allie and Buffler, 1998). This again illustrates the higher level of prior ability that these students possessed.

A superficial perusal of the personal information that the students provided in completing the pre-instruction questionnaire showed that the vast majority of the students attended "advantaged" schools (the term 'advantaged' is used here to describe those schools that are either privately funded or previously fell under the House of Assembly or "white" school system). These schools are known to have functioning science laboratories and are generally staffed by well-trained and equipped teachers. An assumption may thus be made that students in this study have been exposed to various practical experiences and are familiar with many instruments and methodologies used in laboratories.

In light of all of the above, and considering the previous studies on special access students, it is the aim of this work to answer the following research questions:

- To what extent is the model of point and set paradigms useful in interpreting the ideas about measurement held by mainstream students?
- How do the views of these students develop (in terms of point and set paradigms), after completing a full year laboratory curriculum, in the three areas of data collection, data processing and data comparison, and how do they differ between areas?
- What differences exist between mainstream and GEPS students with respect to their understanding of the nature of measurement and uncertainty?

# 2

# Methodology

## 2.1 Research Instrument

### 2.1.1 Instrument selection

The study reported on in this work uses qualitative research methods, as the chief aim of conducting the research was to gain insights into students' understanding. Publications in the fields of research in the education and social sciences (e.g. Cohen *et al.*, 2000; Bell, 1999; and Patton, 1980) discuss the different frameworks for these methods. The present work follows the survey type of investigation, since the information students provided was analysed in order to extract patterns and make comparisons (Bell, 1999). Some disadvantages of this approach for eliciting student ideas are: the issue of representivity, the difficulty of wording questions such that they mean the same to all respondents, and that the choices made by respondents on survey type questionnaires rarely provide sufficient information as to why the individuals acted or reasoned in a particular way. The design of the research instruments used in the studies at UCT attempted to compensate for most of the shortcomings expressed in the literature. Subsequent sections deal with these aspects in more detail.

Data on individuals' ideas and understanding may be collected through questionnaires or by means of conducting one-on-one interviews. However, both methods of collecting data may be employed in any particular research study. The use of the term *questionnaire* in the context of this work is to describe any research instrument for which students are required to either make a choice amongst a few options or provide free responses (open-ended) to posited questions. For self-completion

questionnaires, students are required to complete a questionnaire without the mediation of an interviewer. Of course, questionnaires may include both multiple choice and free response items.

The *interview* may be described as "a conversation between interviewer and respondent with the purpose of eliciting certain information from the respondent" (quotation from Bell, 1999). Various publications draw attention to the specific advantages and disadvantages of collecting data by means of questionnaires and interviews (Bell, 1999; Cohen *et al.*, 2000; Oppenheim, 1992; and Gillham, 2000). Interviews provide a personalisation of students' responses and are adaptable, allowing a skilled interviewer to follow up ideas, probe responses and investigate students' motives and feelings. A large degree of control of the student sample is possible, unless students refuse to take part or miss appointments. Students generally express themselves better verbally, so more complete and detailed data may be collected. There are problems however with using interviews for an investigation of students' ideas. They are time consuming, thus limiting the size of the student sample; are prone to interviewers' subjectivity and bias; involve complex and time-consuming data reduction due to coding demands; and generally produce results that cannot be considered to be reliable overall (see for example Bell, 1999; Cohen *et al.*, 2000).

Questionnaires are relatively inexpensive and places modest time demands on researchers. They may be regarded as reliable due to the relative ease of the analysis process, are easy to construct and make it possible to gather information from large numbers of respondents quickly. An interviewer's personal bias is largely removed from the data-gathering phase, although it must be noted that questionnaire design may introduce personal biases as well, in terms of wording, selection of questions and evaluation methods. These issues are addressed later. That all respondents have to answer the exact same set of questions, removes another source of bias. On the other hand questionnaires that are poorly designed lead to defective information, require reasonable writing skills from respondents (open-ended written questionnaires), and do not allow for misunderstandings of questions to be rectified. There are no opportunities for asking students questions or probing their responses beyond what they have given on the questionnaire, nor to check the seriousness or honesty of their answers. There are also issues of anonymity, the wording of questions, the order of answering questions, and the problems of motivating students to take part in the study. However, careful instrument design and a considered protocol for administration allow most of these concerns to be addressed.

Questionnaires may have different formats. A multiple-choice format allows for quick and easy analysis of questionnaires. However, questions of this format, if not carefully worded, could lead to responses based on biases expressed in the choices provided. When open-ended or free response

questions are used, researchers typically are interested in students' ideas on particular aspects as in this study. These types of questions are however more difficult to analyse and must be limited in form and number to satisfy cost and time constraints (again, see Bell, 1999; Cohen *et al.*, 2000; Gillham, 2000; and Oppenheim, 1992).

### 2.1.2 The semi-structured open-ended questionnaire

As this study formed part of a larger project conducted by the UCT-York group, involving many cohorts of students, it was advantageous to keep the analysis process as simple and unambiguous as possible for logistical reasons. The questions (*probes*) used in the PACKS study (Lubben and Millar, 1996) provided a starting point for the design of the instrument used in the studies at UCT. However, the PACKS probes were designed for school children aged 11-15 based on a range of different contexts. The group recognised that respondents were likely to have difficulties in visualising hypothetically-posed scenarios and required an experimental context relevant to university laboratory work. Consequently, the PACKS probes were considered unsuitable for university students and new probes were designed specifically for the studies at UCT (Allie *et al.*, 1998; Buffler *et al.*, 2001; and Lubben *et al.*, 2001) using a single and easily recognisable context. This addressed many of the potential weaknesses of a survey questionnaire designed to elicit written responses from students, as previously mentioned. The studies mentioned above (Lubben and Millar, 1996; Allie *et al.* 1998) have identified and expanded on the major categories of understanding in the area of experimental evidence, indicating the validity and efficacy of the specific methodology and research instruments used.

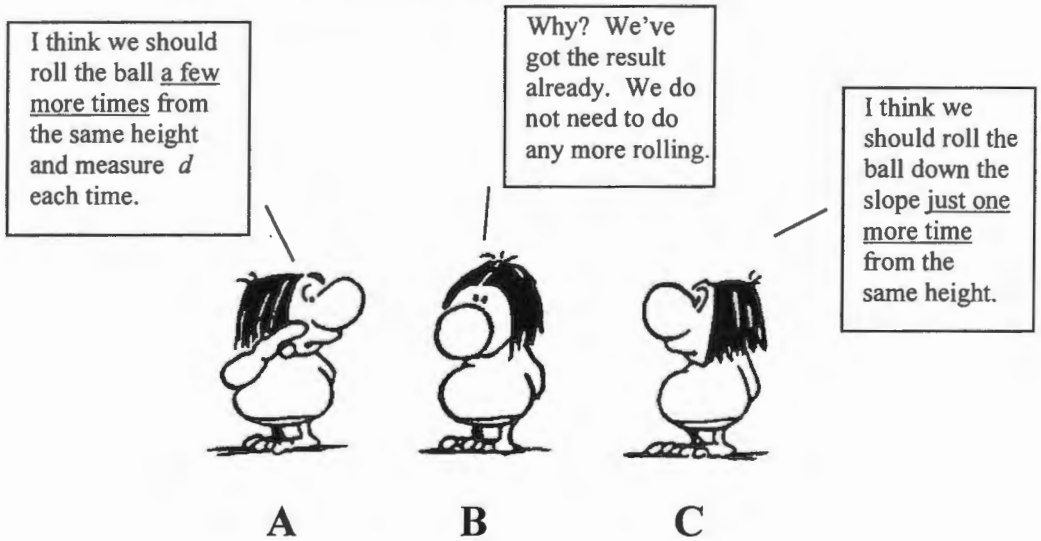
The Allie *et al.* (1998) study used questionnaires and interviews to probe students' ideas on measurement. The analysis of the questionnaires was undertaken without much difficulty and comparisons between responses before and after instruction was found to be very reliable. A sample of students was interviewed to test whether the coding and categorisation of students' responses agreed with the deductions made from the interviews. The interviewers concluded that the reasoning inferred in the written justifications of the students interviewed were a good reflection of their actual reasoning. Only a small percentage of students (less than 10%) indicated that they would have liked to change their responses if they had the opportunity.

Figure 2.1 presents the format of a typical question (the RD *probe*) included in the questionnaires. The probes included in the questionnaire followed the open-ended structure since the study focused on documenting the variety of students' ideas on experimental evidence, whether scientifically correct or incorrect. The multiple-choice format was thus inappropriate for the study. The

majority of the probes may be considered semi-structured since students were required to make a choice amongst two or three options for which they had to provide a written justification. Although Figure 2.1 only shows two lines for students to write their explanations, the actual probe sheets (in Appendix I) had several more lines, encouraging students to write out their answers in full. The choices provided drew students' attention to the aspects of evidence probed thus restricting the possible range of responses and consequently simplifying the analysis process.

The students work in groups on the experiment. Their first task is to determine $d$ when $h = 400$ mm. One group releases the ball down the slope at a height $h = 400$ mm and, using a metre stick, they measure $d$ to be 436 mm.

The following discussion then takes place between the students.




With whom do you most closely agree? (Circle ONE): | A | B | C |

Explain your choice.

_____

_____

*Figure 2.1: The RD ("Repeating Distance") probe.*

The instrument used in the Allie *et al.* (1998) study was administered to students who were largely English second language speakers and lacked real laboratory experience. This required the

questions to have a terse, linguistically uncomplicated writing style. Previous studies (Toh and Woolnough, 1990; Kaunda *et al.*, 1998) have shown that students often lack the terminology to effectively communicate their reasoning on experimental procedures. The individual items were thus structured around scenarios where experimental decisions are debated and the respondents asked to side with one of the suggested courses of action and to then justify their choice by providing a written response. The posited scenarios presented the appropriate terminology without providing procedural hints. In questions where a particular action could indicate a student's reasoning or procedural understanding directly, no suggested courses of action were given (free response) allowing the student maximum control as what to do with presented data or situations.

Although the students in this study were linguistically more advanced (typically first language English speakers) than the students in the Allie *et al.* (1998) study, the style and format of the questions as described above were just as appropriate. This study thus used the instrument as developed for and validated by the Allie *et al.* (1998) study, including further probes developed for later studies (Lubben *et al.*, 2001; Buffler *et al.*, 2001).

### 2.1.3 Description of questionnaire

The research instrument comprised written pencil-and-paper question sheets with one question per sheet (the probes). The questionnaire was administered at the beginning of the academic year before commencement of the mainstream first year laboratory course and a modified questionnaire which contained additional probes was administered after the course. The particular aspects of measurements which the individual probes attempted to investigate will be discussed in detail in Chapter 3. A single posited 'experiment', which provided the sole basis for all the probes, was demonstrated and explained to the students before commencement of the questionnaire. The responses to seven probes were analysed in the pre-instruction study. Three probes dealt with the reasons for repeating measurements, which addressed the area of data collection. Two probes investigated the decisions students made in handling data sets both analytically and graphically, which are aspects of data processing. Two probes dealt with the quality and compatibility of data sets, which required respondents to make judgements based on two different sets of measurements provided in the probes. These questions probed aspects of data comparison.

Eight questions from the post-instruction questionnaire were analysed for this study. Two of the data collection probes used in the pre-instruction test were included. The data processing probes were used unchanged. The data comparison probes were used as per the questionnaire before instruction with two additional probes. One presents the results of five repeated distance

measurements together with a calculated mean value (similar to the first two), the other provides the measurement results in terms of means and standard deviations of the means.

### 2.1.4 Experimental setting

A First Year Physics Laboratory experiment, used in the Physics Department at UCT as a laboratory examination, provided the context for the probes. The students would thus not have encountered the experiment either before or during their course which was important for avoiding the effect that prior knowledge and experience could have had on probe responses. The choice of experiment would seek to limit rote responses to the questions and lessen student misconceptions due to its simplicity and straightforward description. The experimental context for this study is identical to the previous studies on special access students (e.g. Allie *et al.*, 1998). The interviews that validated the research instrument in the Allie *et al.* study confirmed that students understood the questions and provided answers that explained their reasoning faithfully. The posited experimental setting would then likely not be problematic for the more advanced cohort of students in the present study.

The actual laboratory experiment requires the following apparatus: a small sloping wooden ramp, a clamp, a steel ball, a metre stick and marking paper. The experiment, as demonstrated to the students before commencement of the questionnaire, employed a large-scale wooden slope and a tennis ball instead of the small ramp and steel ball. The apparatus needed to be visible from some distance in the large lecture theatre used for administering the questionnaire. The stem of text (or text statement) and the diagram in Figure 2.2 describes the experimental setting for all the probe questions, and appeared on the front of the envelopes in which the individual probe sheets were placed.

### 2.1.5 Demonstration apparatus

The Physics workshop at UCT constructed the scaled-up model of the apparatus used in the first year laboratory experiment. A brightly coloured tennis ball replaced the steel ball. Figure 2.3 contains a photograph and indicates the dimensions of the model.

The wooden slope tapers to a horizontal ramp, which ensured that the ball was launched horizontally in the demonstration of the experiment. The diagram of the setup, shown in Figure 2.2 and made available to the students on the cover of the questionnaire envelope, clearly indicated the horizontal projection of the ball. The demonstration needed to be clearly visible to all the students

taking part in the questionnaire. The tennis ball was released from two different heights to demonstrate how the distance, *d*, on the floor changes as the height, *h*, is varied.

An experiment is being performed by students in the Physics Laboratory.
A wooden slope is clamped near the edge of a table. A ball is released from a height *h* above the table as shown in the diagram. The ball leaves the slope horizontally and lands on the floor a distance *d* from the edge of the table. Special paper is placed on the floor on which the ball makes a small mark when it lands.

The students have been asked to investigate how the distance *d* on the floor changes when the height *h* is varied. A metre stick is used to measure *d* and *h*.

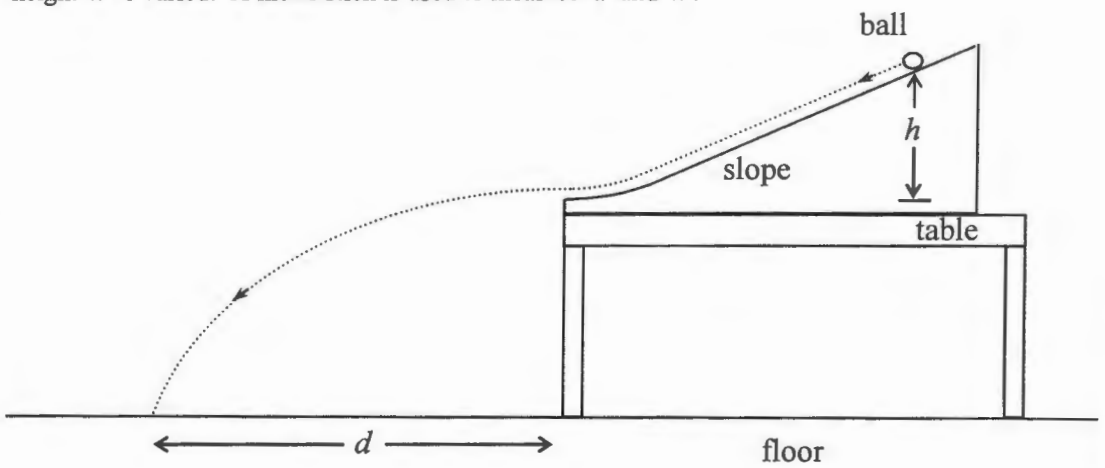



*Figure 2.2: Description of experiment and diagram of experimental set-up. This provided the basis for the procedural questions in all the probes.*

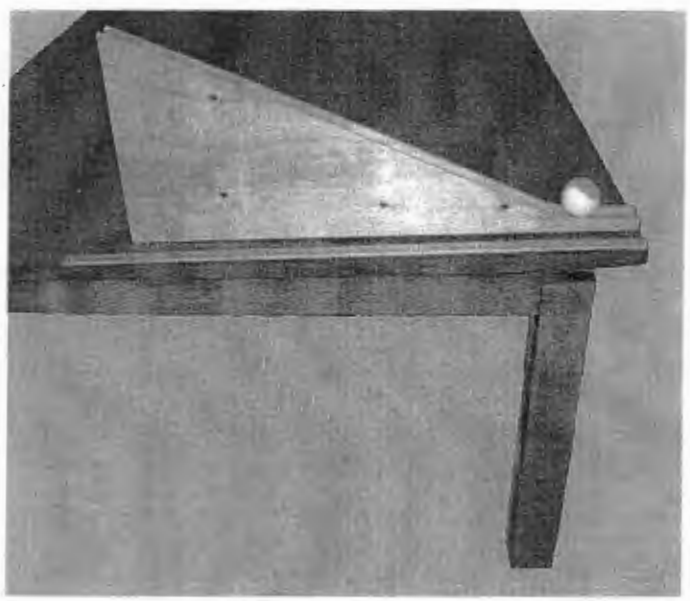


*Figure 2.3: The wooden ramp and tennis ball used to demonstrate the experiment. The dimensions of the ramp are 40 cm high, 6 cm wide, 90 cm long; the metre stick provides a length perspective.*

### 2.1.6 Instruction sheet

The instruction sheet was pasted onto the front of a large A4-sized brown envelope, which enclosed the individual probe sheets. The instruction sheet appears in full in Appendix I. Each student wrote his or her surname and first name in the block at the top of the sheet. The letter in the block below and to the left of the name block differentiated between the versions of the questionnaire. The letters D and E denoted the questionnaires used before and after instruction respectively. This also provided a convenient means of differentiating between the various sets of probes used in different studies in the larger research project. A unique numerical number was stamped in the block to the right and below the name block. This number was also stamped onto every probe sheet included in the questionnaire. This ensured that each probe sheet could be uniquely identifiable with an individual student but additionally, the responses to the probes would remain anonymous to the person analysing them. The numbering of the probe sheets also facilitated the analysis process as different probe responses were compared for individual students. The lines of text below the name block identified the institution and department within which the study was undertaken and stated the instrument title – "Laboratory Procedures Questionnaire".

An instruction list enclosed in a text box then followed. The motivation for placing the instructions first and in a box was to focus the students' attention on the protocol details before engaging the experimental details. The context and diagram of the experimental set-up (see Figure 2.2) was placed next on the sheet. The diagram augmented the stem of text (under the heading Context). A detailed scaled drawing of the apparatus was purposefully avoided so that the students would focus on the text when read to them before answering the questionnaire.

By placing the instruction sheet on the envelope cover, students could refer back to the instructions and the description of the experiment continuously when answering the probes.

### 2.1.7 Probe sheets

The A4-sized probe sheets were placed in numerical order into the large brown envelopes. Each probe sheet contained one probe question. Although a largely automated process was used to print and collate the sets of sheets, the researcher manually checked each individual set to ensure that no sheet was omitted due to some technical hitch and that each set was indeed in the correct ascending numerical order (Question 1 on top, followed by Question 2, etc.) before insertion into the envelopes.

The design and appearance of each probe sheet followed a similar format. The question number with a letter code in brackets (e.g. RD) appeared at the top of the sheet. This was to draw the students' attention to the numbering of the probes, which would help with adherence to the protocol instructions. The letter code, in capital letters, loosely abbreviated the description of the probe, e.g. RD is the letter code for the "Repeating Distance" measurements probe. Immediately below the question number and abbreviated probe description (the version of the probe was also included), a stem of text introduced to respondents an imagined scenario where a group or groups of students would be performing a task or procedure related to the posited experiment. This formed the basis of the question. A discussion setting followed (with the exception of the data processing probes, UR and SLG, details later) where different imaginary students or groups would be engaged in a discourse about the course of action that should be followed, reflecting the various procedural options that could conceivably be taken. Clearly labelled cartoon characters represented the students. The specific cartoon characters (from "King Tut" by Geoff Watson) were chosen to avoid any gender, cultural or race biases that may have negatively influenced students' responses. Quotation blocks enclosed the characters' comments. The probe then required of respondents a choice that reflected their procedural decision by siding with one of the cartoon characters. The students indicated this by circling a letter (A, B, C, etc.) in the multiple-choice block below the cartoon characters. Finally and most significantly, the students were requested to provide a detailed explanation of their choice. It was critical for students to communicate their ideas regarding their procedural decisions clearly and fully, hence a sizeable fraction of the page area, more than a third typically, was reserved for the explanation. The horizontal lines separated consecutive lines of text, which aided the reading of the responses during the coding process.

In both the pre- and post-instruction questionnaires, the last probe afforded students the opportunity to indicate whether they would have liked to change the answers or responses to any of the probes and make additional comments.

## 2.2 Protocol

Strict protocol rules governed the administering of the questionnaire. Most of the books written on research methods in education dedicate sections to issues such as anonymity, confidentiality, disclosure of the purposes of a study and the use of the collected data (e.g. Bell, 1999; Gillham, 2000; Cohen *et al.* 2000; and Patton, 1980). The research instrument was designed to address these concerns. Students completed the questionnaires in an examination setting, enabling the researchers to explicitly deal with the protocol issues before introducing the details of the questionnaire.

### 2.2.1 Confidentiality and context

The students completed the pre-instruction questionnaire in the first week of lectures at the beginning of their academic year. Students would thus have received course handouts and evaluation regimes by the time they wrote the questionnaire. It was therefore deemed important to explain to students that the results of the questionnaire would not have any influence on their course records. It was also made expressly clear that participation in the study was entirely voluntary; the students were given an opportunity to decline and leave if they chose not to take part in the study. It was compulsory for the students to write their names in the space provided on the front cover sheet. The researchers explained to the students that this information would only be used to generate an anonymous profile of the student cohort. This was also the only way to link individual students to their pre- and post-instruction responses to the probes. The researcher that analysed the data did not have access to class records or official university lists so no impact was possible on course performance and ensured the anonymity of the results. The researchers stressed that the questionnaires would be treated with total confidentiality and would in no way bias any aspect of the students' course evaluation. The students were informed that the purpose of the questionnaire was specifically for research and that no other use would be made of the data. The questionnaire administered after instruction followed the same protocol.

### 2.2.2 Instruction and demonstration

Each student taking part in the questionnaire received a brown envelope with the instruction sheet on the front cover and the probe sheets inside. By separating the instructions from the probe sheets in this way, the students were forced to first pay attention to the instructions and context before attempting to answer the probe questions inside the envelopes. The instructions were read out loud by one researcher in the order in which they appeared on the instruction list. The researcher reading the instructions paused if an instruction required an activity to be completed by the students, e.g. the researcher paused and remained silent while the students completed writing their names in the names block after reading the instruction, "Write your name in the box above". After reading through the instructions once, the researcher repeated the process emphasising the instructions printed in bold lettering.

The first instruction on the list as discussed above lead to students writing their full names in the space provided and a discussion on the issue of confidentiality. Listed second was the comment that appeared immediately below the first instruction, "Inside this envelope there are pages

numbered up to page n." Students could then check that their questionnaires included all the probe sheets in the correct order. Third followed an instruction, "Read the text below and answer the questions on each sheet." Naturally, it was important that students understood clearly what was required of them and what the questionnaire was all about. Fourthly, a comment was included, "If you need more space for your answers, then use the backs of the sheets." The probe questions were printed on a single side of white A4-sheets. Students might have needed more space to complete their answers and were encouraged to complete their responses on the back of the particular probe sheets. A comment, "It should take you about 5 minutes to answer each question", appeared fifth in the list. This comment set a loose time margin for answering the questionnaire. A time of five minutes was deemed adequate to answer each question. The time students spent on answering the probes was an important consideration, as too little time would result in unsatisfactory consideration of the issues.

The text in bold lettering focussed the students' attention on the importance of those protocol details. The instruction "**Answer the questions in order and do not skip any sheet**" was included to emphasise the importance of students answering the probes in strict sequence. This was required since subsequent (following) probes could hint at "correct" responses or bias a student's reasoning in preceding probes. The probe questions were designed with this protocol in mind, ordering the probes to minimise the effect of procedural hinting. An example of this would be that probes that explicitly referred to an average (or mean) of readings would in all likelihood bias a student's responses when answering probes which deal with decisions regarding repeating of measurements. Consequently, probes that explicitly refer to a mean were placed later in the questionnaire. The instruction, "**When you have completed a question, put the sheet inside this envelope and do not take it out again, even if you want to change your answer",** was included to discourage students revising their responses to particular probes. Subsequent probes may have prompted students to rethink their responses to completed probes. The most important information is the ideas evidenced in the students' written responses, hence allowing students to change their answers would have 'contaminated' responses. The last item addressed this, "**Note: It is possible that some answers may be similar or exactly the same as others. Please write all answers out in full, even if you feel that you are repeating yourself."** Students might have been inclined to abridge certain responses if they believed that answers given for preceding probes could explain their decisions for subsequent ones. The probes were coded probe-by-probe and not by student, requiring full explanations for each probe.

At this point the students were reminded that an opportunity existed for them to indicate whether they would have liked to change any of their probe responses on a final probe sheet at the end of

the questionnaire. Respondents would not be particularly concerned with the "correctness" of a response if they knew that there was an opportunity to indicate a change of mind or modification of an answer at the end.

The text referring to the posited experiment was then read out loud by one researcher while another researcher demonstrated the experiment by releasing the ball from two different heights; chosen so that the students could observe the influence that the height had on the distance the ball travelled before landing on the floor below. The full text for the *Context* appears at the bottom of the instruction sheet on the front cover of the brown envelopes (Figure 2.2 and Appendix I). The students were instructed to observe the procedure. The demonstration apparatus was positioned in the test venue to ensure that every aspect of the demonstration (the slope, ball being released, ball landing on the floor, etc.) was clearly visible to all the students. The researchers repeated the reading of the text and the demonstration of the procedure one more time, in an identical manner as the first time. To ensure that students received no procedural hints, the researchers rehearsed the demonstration prior to administering the questionnaire and exercised care in conducting themselves in a neutral, subdued and unanimated manner.

## 2.3 Analysis methodology

On completion of the questionnaires, students handed over their brown envelopes with the full set of probe sheets inside. It was the task of a researcher to separate out individual probe sheets and collect into sets the responses of all the students to individual probes, and sorted according to set number. As described in a preceding section, the unique set numbers identified individual students and facilitated the spreadsheet analyses that followed. The sorted sets of responses facilitated the coding process that followed, as it was easy to page through individual responses whilst keeping the sets neatly together.

### 2.3.1 Coding of probe responses

It was realised from the outset of the project that an analysis process premised on the use of keywords (accuracy, precision, best-fit, etc.) in written responses would not be appropriate for gaining deeper insights into students underlying understanding of measurement. As noted previously and supported by the literature (e.g. Garrett *et al.*, 2000; Séré *et al.*, 1993; and Evangelinos, 1998), students both at school and university level use terminology associated with experimental work in haphazard, confusing and/or erroneous ways. In addition to these

considerations, a large part of the research programme was undertaken with English second language users. These factors suggested a more careful approach for categorising written responses. The research sought to investigate students' ideas about measurement to ultimately gain knowledge of their understanding when performing actions and procedures in the physics laboratory.

The first step of the analysis process involved the development of coding schemes for each probe. The choice of action (A, B, C, etc.) together with the written explanation formed the basis of the codes. The Grounded Theory method (Strauss and Corbin, 1990) was followed in developing the categories of responses to individual probes. Individual responses were systematically considered in generating the categories by identifying key ideas and grouping responses. As students' actual responses provided the descriptors for the categories, the process involved clarifying and refining descriptors and in some instances required amendment to better represent response types. As it was desirable to differentiate as many unique ideas expressed in the responses as possible, and since the goal of the exercise was to identify the students' reasoning, categories were subdivided and delineated to make them mutually exclusive where necessary. This resulted in a draft of the coding scheme for each probe, which was used to code sets of responses independently. The assigned codes were compared with the descriptors, which were refined as required to produce a valid coding scheme. Different research team members used the schemes to code responses independently, thus verifying the coding schemes. Differences in codes assigned (less than 10% of the time) were resolved by inspecting responses from an individual student across clusters of related probes.

For this study, many of the probes had previously developed coding schemes, used for the earlier studies and cohorts. These schemes generally were found not to encompass the responses from the linguistically more advanced mainstream students in this study. There was a greater level of sophistication in these students' answers in terms of both language usage and procedural reasoning. The existing schemes were thus extensively revised and updated for this study. Newly designed probes (e.g. DMSU) required that schemes be developed following the entire process. The coding schemes developed for each of the probes used in this study will be discussed further in Chapter 3.

Another feature of the analysis process employed for this study was to test how well students' responses fit the analysis framework developed earlier (Allie *et al.*, 2001), in terms of the point and set paradigms. The point and set paradigmatic model was used previously (Buffler *et al.*, 2001) to analyse the responses of special access (GEPS) students. For this study, every descriptor representing a response category for all the probes were compared to the definitions of point and

set paradigms (Tables 1.2 and 1.3) and assigning paradigm codes to every descriptor (P for point, S for set, U for undecided). Ambiguity and contradictions were resolved by looking at an individual student's responses across related probes within the different areas of measurement. The results of this procedure will be presented in detail in Chapter 3.

### 2.3.2 Cross probe analyses

The coding schemes were structured to enable the underlying reasoning to be identified for each student across the different measurement-related situations, namely data collection, data processing and data comparison. Similar reasoning could be used in different probes where different actions were taken. Individual students' responses to all the probes within and across the three areas of measurement were classified within the paradigmatic model by considering all the probes together. In other words, the point and set codes, which were assigned to the code category descriptors, were not the sole determinants of a students' classification in terms of the point or set paradigms within a particular area of measurement. This then also provided a measure for determining the validity and efficacy of assigning paradigm codes directly to response category descriptors. Frequency tables were drawn up for each probe, both before and after instruction with tables comparing pre- and post-instruction use of paradigms within and across the three areas of measurement.

# 3

# Analysis of questionnaire probes

The point and set paradigmatic model that was developed and tested by the UCT-York research collaboration was described in some detail in Chapter 1. This model provided the analysis framework used in this study. This chapter describes in detail the individual probes in the areas of data collection, data processing and data set comparison. Coding of student responses was carried out as described in Chapter 2. The rationale of the coding scheme and the code assignment regime will be presented in detail. Categories of ideas underlying the students' responses will be highlighted and justified. Actual student responses that typify response categories will be quoted. The development of the coding schemes encompassed responses from a diverse group of students. However, only responses from the cohort of mainstream students, the focus of this dissertation, will be quoted.

## 3.1 Instrument Questionnaire

The individual probes investigated different elements of understanding of measurement, uncertainty and experimental evidence, and dealt specifically with aspects of data collection, data processing or data comparison. Appendix I contains the complete set of probes used in this study.

### 3.1.1 The data collection probes

The data collection probes used were RD (*repeating distance*), RDA (*repeating distance again*) and RT (*repeating time*). These probes were designed to investigate students' reasons for

collecting data through repeating measurements of the same quantity, or their reasons for electing not to repeat readings. Chapter 2 highlighted the importance of the order in which the probes should be answered. The pre-test required the RD probe to be answered first, RDA second and RT seventh. The RT probe appeared later to test whether the students' actions and reasoning when collecting data would be influenced by a different context (time versus distance). By the seventh probe, students would have seen other probes where several measurements were taken with a couple explicitly mentioning a mean value. The possible effect of other probes on responses to the RT probe is discussed later. The RT probe tested context dependency with respect to ideas about measurement. Only the RD and RDA probes were included in the post-instruction questionnaire, answered first and second respectively.

The RD probe (*repeating distance*) investigates the students' most basic ideas about whether repeating measurements is necessary in experiments and for what purposes multiple readings are needed, if at all. The probe presents the result of a single measurement of the horizontal distance $d$ travelled by the ball from the edge of the ramp before hitting the floor. A situational discussion, that takes place between students of an imaginary group, then provides three alternative procedural decisions; namely to repeat several times, repeat once only, or not to repeat at all. The respondents are required to choose with which student they most closely agree and to give a written justification for their choice. The RDA probe (*repeating distance again*) confronts the respondent with an additional, different result for the same measurement (ball released from the same height $h$). Three procedural decisions, similar to those in RD, are introduced by means of a similar situational discussion. This probe then explores how the students reason when obtaining slightly different results. The students may change their procedural decisions and/or modify the justification for their choice. The RT probe (*repeating time*) is identical to the RD probe except that the measurand is time instead of distance.

### 3.1.2 The data processing probes

The data processing probes are UR (*using repeats*) and SLG (*straight line graph*). The design of these probes allows for the testing of students' handling of experimental data. The students' ideas about, and their reasons for executing data processing procedures they employ in the physics laboratory are explored. The UR probe requires students to represent a set of data, the results of five releases of the ball ($d$) from the same height $h$, with a single quantity. This probe does not present the respondents with any options so no particular procedure is suggested, e.g. to calculate a mean. The SLG probe requires graphical modelling of a series of plotted points which show a trend towards a straight line. The plotted points are carefully arranged to be consistent with real

data obtained from an actual experiment. The points are scattered and positioned so that a wide range of possible lines could be drawn to model the data, e.g. some points are aligned with each other, others with the origin. The order of these probes was identical for both the questionnaires prior to and after instruction. The UR probe was answered after the data collection probes but before the data set comparison probes where the mean is explicitly used in the questions. Preceding probes would thus not have given any procedural hints, e.g. to calculate a mean. The SLG probe was the last probe to be answered in both cases.

Analysis of the AN (dealing with an anomalous measurement result) probe, included in the pre-course questionnaire, was excluded for this study as the results do not answer the main research questions. This probe was placed fourth in the order, after the RD, RDA and UR probes, but before the data comparison probes (SMDS and DMSS) and the RT and SLG probes. Students' responses to subsequent probes thus needed to be considered in the light of exposure to a data set that included a clearly anomalous result. Students' view of variation and spread in results may have been influenced in terms of point and set reasoning, as some actual responses to later probes indeed indicated.

### 3.1.3 The data set comparison probes

The probes that deal with the comparison of data sets are SMDS (*same mean different spread*), DMSS (*different mean same spread*), DMOS (*different mean overlapping spread*) and DMSU (*different mean same uncertainty*). These probes require students to comment either on the compatibility or the relative quality of two data sets. The SMDS, DMSS and DMOS probes present two sets of measurements (five values of $d$) with their calculated means, whereas the DMSU probe presents the measurement results of two student groups in the formal manner, i.e. a mean and the standard deviation of the mean. The SMDS probe introduces the concept of spread explicitly in one of the options. This probe investigates whether students can recognize the spread in data as an indicator of the relative quality of two data sets. In general, the students' interpretation of spread is explored. The DMSS, DMOS and DMSU probes require students to decide on the compatibility of two data sets by determining whether or not the uncertainty intervals (standard deviation), the formal construct indicating the measure of spread, of the data sets overlap. The means of the data sets in the DMSS and DMOS probes are different. In the DMSS probe, the mean of one set falls within the uncertainty interval of the other with ranges of equal size (difference between highest and lowest value). In the DMOS probe, the means fall outside each other's uncertainty interval but their ranges overlap somewhat (although the two intervals defined by one standard uncertainty do not overlap). The DMSU probe presents the results of two

measurements with calculated means, which differ, and one standard uncertainty (standard deviation of the mean), which is the same. Here, the intervals of the results overlap but the means fall outside each other's interval. Only the SMDS and DMSS probes were included in the pre-test. They were positioned in order after the UR probe and before the SLG probe. The DMOS and DMSU probes were added in order after the SMDS and DMSS probes in the post instruction questionnaire.

## 3.2 Alpha-numeric coding scheme

The rationale of the coding scheme used for this study is based on that developed for a previous study of special access students (Allie *et al.* 1998). An analysis framework centred on point and set paradigms was developed to investigate undergraduate physics students' ideas on measurement. The student profile for the current study is more advanced than the cohort described in the Allie *et al.* study as explained in Chapter 1. It was expected that the student responses would be more sophisticated than those of the special access students. Therefore the coding scheme was expanded and for several probes considerably modified. New coding schemes for two probes (DMOS and DMSU) were developed following the Grounded Theory method as explained in Chapter 2.

A feature employed in this study was the assignment of paradigm codes to identify individual student responses, and hence response categories, within the point and set paradigm scheme where possible. A capital letter P was used to classify a student's response as consistent with the point paradigm and a capital S for consistency with the set paradigm. In the few cases that the response could not be classified unambiguously as point or set, a capital U (undecided) was assigned. The wording of these responses was generally characterized by vagueness and ambiguity but not lack of clarity, as was the case with many responses from special access students in a previous study (Buffler *et al.* 2001). It is believed that these students were simply negligent with their responses, since the cross-probe analysis showed that most of these students were easily identified within the point and set model when considering probes together, not individually.

After assigning the capital letter representing the paradigm code, a capital letter (A, B, C, etc.) was assigned representing the choice made by the student. In the event that no choice was indicated, an attempt was made to infer the choice from the written response; otherwise the response was coded N (no response) or U (uncodeable). Certain probes allow open-ended responses where no choice is available, namely UR and SLG. In this case the letter was assigned according to response type (e.g. for the UR probe, using all the readings to calculate the mean was assigned an A).

Apart from the SLG probe, a two digit number code was assigned to indicate the ideas about measurement reflected in the response. The number code was chosen to highlight response patterns, e.g. all responses that explicitly referred to a mean or average were assigned a number code starting with a 2 (20, 21, etc.). The second number allowed for further categorization and again was chosen to indicate ideas about measurement, e.g. the number 24 was assigned to responses where a mean was calculated to "get closer to the true value", and 25 for responses where a mean was calculated to increase accuracy, "be more accurate". Frequently the same number codes were used for the various choices (A, B or C). The coding framework thus accounts for the possibility that similar ideas about measurement could lead to different procedural actions.

The alpha-numeric code was written directly on the individual response sheets for easy data capture into spreadsheets for later reference, analysis and comparison. The analysis process was facilitated by keeping the letter (paradigm and choice) and number codes (reason for choice) together. Analyses could be undertaken to identify major response trends and to study individual students' consistency of use of paradigms and reasoning across probes.

## 3.3 Coding of student ideas

This section summarises the coding process by presenting and justifying the coding schemes used for each probe. Students' written responses will be quoted in order to illustrate the response categories. Minor 'corrections' (capital letters, full-stops, etc.) to quoted text are made only in instances where it is necessary for the readability of the response. The coding was undertaken using the alpha-numeric scheme described in the previous section. The step-by-step details of the development of the schemes will not be presented (the grounded theory procedure mentioned before). However the final coding scheme for each probe will be shown and explained in the following subsections. It must be noted again that the coding scheme encompasses all responses from all cohorts in the larger study (special access and mainstream), hence not all the codes in the schemes were assigned in this study. Response quotes from this cohort were thus not available to illustrate each and every code category. The inclusion of all the code categories serves to demonstrate the coherence and rationale of the schemes.

In the subsections below, the text as used on the actual response sheets will be presented to introduce the discussion on the procedural choices and accompanying reasoning represented by the

code categories. The cartoon characters (as shown in Chapter 2) and spaces for written answers will be omitted. The full-page response sheets are available for inspection in Appendix I.

### 3.3.1 Repeating distance measurement probe (RD)

The RD probe solicits students' ideas concerning the purposes of doing repeat measurements (distance). The context of the probe, the situational discussion and the three procedural options available are presented below:

> The students work in groups on the experiment. Their first task is to determine $d$ when $h = 400$ mm. One group releases the ball down the slope at a height $h = 400$ mm and, using a metre stick, they measure $d$ to be 436 mm. The following discussion then takes place between the students:
>
> **A**: I think we should roll the ball a few more times from the same height and measure $d$ each time.
>
> **B**: Why? We've got the result already. We do not need to do any more rolling.
>
> **C**: I think we should roll the ball down the slope just one more time from the same height.

The coding scheme is presented below:

*UN00* - *No response*
*UU00* - *Uncodeable*

***I agree with A because ...***
*UA00* - *No reason*
*UA01* - *Uncodeable reason*
*PA10* - *You need to practice and then take one correct/valid reading of d*
*PA11* - *Practice to minimize/ take into account outside factors to find correct/ valid d*
*PA12* - *Practice to eliminate "errors"/ mistakes/ discrepancies to find correct/ valid d*
*PA15* - *Practice will allow you to get the accurate measurement of d*
*SA20* - *More measurements are needed to get an (better) average/ mean*
*SA21* - *Get average/ mean to reduce effect of outside/ random factors*
*SA22* - *Get average/ mean to reduce effect of errors/ mistakes*
*SA23* - *Get average/ mean and (better/narrower) spread/ uncertainty/ standard deviation*
*SA24* - *To get an average/ mean to get closer to the true value*
*SA25* - *Get an average/ mean to be more accurate/ get a more accurate answer*
*SA26* - *Get average/ mean that is an estimate/ approximation of the distance d*
*PA30* - *Get the recurring/ same/ correct answer - be confident of answer*
*UA40* - *Ensure that d does not vary too much with each release (actual d related to h)*
*UA41* - *Check that outside/ random factors did not influence results*
*UA42* - *Experimental error in d is reduced/make sure that expt'al error is not too big*
*UA43* - *Eliminate/ reduce random errors in readings (to be exact)*
*UA45* - *To obtain accurate measurements for the distance traveled d*
*UA62* - *Reduce effect of errors on result*
*UA65* - *Get a more accurate answer/ measure of d*
*SA70* - *To gauge/ determine the spread/ uncertainty in the measurement*
*SA73* - *To determine a better/ narrower spread/ uncertainty*
*SA75* - *Inherent experimental uncertainty in any experiment→ repeat to get accurate representation*
*SA81* - *Decide to take recurring value or an average if variation*
*SA82* - *Gauge variation in d and calculate mean to reduce effects of experimental error, decide on more repeats if difference in d significantly large*

***I agree with B because ...***

*PB00* - *No reason*
*PB01* - *Uncodeable*
*PB30* - *Repeats will give the same result (h unchanged, hence d unchanged)*
*PB31* - *Repeats will give the same result if outside factors are constant*
*PB40* - *Repeats will give different results which is confusing/ pointless*
*PB50* - *Repeats are a waste of time/ resources*

***I agree with C because ...***

*UC00* - *No reason*
*UC01* - *Uncodeable*
*PC30* - *Check/ confirm the result (get recurring/ same/ correct answer)*
*PC31* - *Confirm/ verify result; all experimental conditions ("the physics") same/ affected by external factors*
*PC32* - *Check that no errors made in obtaining measurement/ confirm validity of results*
*PC35* - *Verify/ check the accuracy/ precision of the measurement*
*PC38* - *Assume reading to be correct if next release the same; if it differs then more repeats*
*UC40* - *Need to repeat experiment more than (at least) once*
*UC42* - *Not many errors expected, rough confirmation of results*
*UC43* - *To test for differences and if any learn more about experiment*
*PC50* - *Many repeats are a waste of time/ resources*
*PC60* - *Choose the correct answer* [Note: no guideline given as how to make choice]
*UC80* - *Decide to take more measurements if variation large*
*SC81* - *Calculate average from two if 2nd result similar; decide to take more measurements (third) if dev large and calculate average*
*SC82* - *Control or confirmation of result - if deviation large, average from many (at least three) measurements*

*An arrow (→) in any of the code descriptors indicates an argument where reasoning is explicitly tied to a procedural action.*

*Code qualifiers:*

*a* := *accuracy* [A student who explicitly referred to the term accuracy]
*p* := *precision* [A student who explicitly referred to the term precision]
*e* := *error/mistakes* [Reference made to experimental error, random error or mistakes]
*x* := *external factors* [Comment on outside influences, external factors or conditions]
*b* := *true value* [Mention made of a true value]
*d* := *deviation* [Argument explicitly noted the deviation (size of) in readings]
*s* := *stopwatch/timing* [Explicit reference to timing procedure]
*t* := *time factor* [Concern expressed over time constraints]
$\sigma$ := *standard deviation mentioned explicitly*
*m* := *standard deviation of the mean mentioned explicitly*

The above code qualifiers were assigned when students gave additional secondary reasons in defence of their choices. Some individuals gave many secondary reasons, in which case more than one code qualifier was assigned.

Respondents who subscribed to the procedural decision of repeating the measurement several times agreed with the suggestion made by cartoon character A by circling the letter. The scheme of codes shown above details the different reasons given by the respondents for siding with A.

*Choice A: Incomplete responses*

When the space provided for an explanation was left blank, UA00 was assigned. The paradigm code (first letter) is U since choosing A does not necessarily imply that the respondent based this choice on reasoning rooted in the point or set paradigm. UA01 was assigned for written responses that could not be classified within the framework. All probes that were not answered, incomplete or that could not be categorised within the framework followed the same assignment regime as described above.

*Choice A: Point paradigm*

Students might want to roll the ball several times to practice and perfect their experimental technique. This is indicative of the point paradigm as it reflects a belief that repetition will eventually yield the perfect or correct measurement result. The codes PA10-15 would have been assigned to such responses. None of the students in this cohort indicated that they wished to practice for several releases and eventually take a final ideal result. However, this type of response was prevalent for the special access students in a previous study (Buffler *et. al.* 2001).

Students could express the need to repeat measurements in order to either check the validity of the result or identify a recurring reading. In this instance, responses would typically indicate the need to be confident of the reading. This type of reasoning shows a firm conviction that a single reading may adequately represent a measurement result, even with scatter present in the data. These responses are typical of the point paradigm and hence coded PA30. One student gave the following written response:

> *One needs multiple results to make sure the measurement is correct.* (PA30)

*Choice A: Set paradigm*

The following code grouping, SA20-26, focus on the requirement of several repeats in order to calculate a mean value (students mostly use the word average here) to best represent the outcome or result of the measurement. Averaging is typically seen as a way of dealing with variation in experimental readings (SA20). This kind of reasoning is compatible with the set paradigm since there is an appreciation for the modelling of an ensemble of data. Responses may also identify external factors (SA21) or experimental errors (SA22) as reasons for variation in readings, or explicitly note the need to establish a mean and a standard deviation (SA23).

*The measurements will probably be different every time you release the ball from the same height because you can never do exactly the same experiment twice. It is better to take an average of a few measurements than only one or two.* (SA20)

*There are numerous factors that could affect your experiment such as air resistance, friction etc. To reduce the effect of these factors you should take an average distance travelled, by dropping the ball a few more times, thus creating a more realistic answer.* (SA21)

*There will be some random error in this experiment, as always, due mostly to small errors made in measuring, and hence taking a single measurement gives just a single point in a distribution. As many measurements as possible should be taken, random errors cancelling to a better and better degree when the mean is taken.* (SA22)

*Taking more readings and finding the average reduces the percentage error. Also, a reading which obviously does not agree with the rest of the readings can be eliminated (due to some error conducting the experiment).* (SA22)

*Many experiments should be performed so that a reliable average and standard deviation can be calculated. This cannot be done with only one or two readings.* (SA23)

A student could opt for averaging to better represent or approach a true value. The conviction about the existence and therefore pursuit of a true value is problematic in terms of a correct understanding of experimentation but nevertheless, these responses are coded SA24 as those given below:

*If you take a number of results it is possible to find an average and therefore get a result closer to the true value.* (SA24)

*The bigger the sample, the more likely the mean will be closer to the real (ideal) value.* (SA24)

Responses that subscribe to the idea that repeating and averaging would improve the accuracy of the result (SA25) or approximate the measurand (SA26), are illustrated below:

*The results are likely to be more accurate if the mean of a number of results are taken. The more trials we do, the greater the accuracy, since our measurements may not be perfectly accurate.* (SA25)

*If the experiment is done several times it is possible for the experimenter to estimate or approximate the distance d. This is done by taking every value of d and add them, look for the average which will be the approximated d value.* (SA26)

In instances where multiple reasons were given for the deviation in readings and hence need to determine the mean value, code qualifiers, shown above in the coding scheme, were used to mark these responses. The assignment of qualifiers applies to all code categories for most probes. A few examples with assigned codes follow:

*Experimental results are never completely accurate, mistakes could be made with measurements, or outside factors like friction or airflow could interfere. The more times the experiment is repeated the more accurate the average result would be.* (SA21ea)

*Because experimental error plays a part in it and one will need to find a mean value from a few values depending on the nature of the equipment (precision & accuracy).* (SA22pa)

*The more readings you have at a height, the better the distribution of possible results will be, and the more accurate the mean result calculated and its standard deviation (i.e. by repeating this, more accurate results can be obtained in order to predict future behavior of the ball).* (SA23a)

*One should always take as many measurements as possible and then calculate a mean value and associated standard deviation. In this experiment, it is particularly important to take many readings as the trajectory of the ball can be influenced by many things.* (SA23x)

The following category of responses does not include the determination of the mean in the argument for several repeats. These responses highlight the need to gauge the spread in experimental data, either by looking at the size of intervals or the uncertainty. The set paradigm clearly underlies statements that recognize spread as an integral part of a measurement result. The numeric part of the codes with qualifiers displays the same pattern as described previously. A few sample responses assigned SA70 to SA75 are presented below:

*When rolling the ball there may have occurred a circumstance that hinders the accuracy of the result of the reading. For example the ball may have been slightly pushed when released or it may have slid rather than rolled. By doing the experiment "a few more times" one can get an idea where the general region of the result should be. Should the experiment be repeated only once again, and two drastically different results obtained, which one would you take us correct? You'd have to repeat it a few more times and discard any obvious error reading.* (SA70xa)

*The more measurement taken, the more accurate values will be, and uncertainties will get smaller.* (SA73a)

The following category, coded within the set paradigm, contains those answers where a decision is made based on the deviation in readings from several repeats. Responses where a mean would be calculated only if a spread in data is observed were coded SA81. Those where a decision, about how many additional repeats are required to calculate a mean, would be made based on the degree of spread were coded SA82. This reasoning is what seasoned experimenters would use when carrying out laboratory procedures; hence this kind of response is very advanced for entry-level students. The sample responses below were coded with the numeric and qualifier code assignment pattern as before:

> *Each time you take the measurement you have a better idea of the accuracy. If the results are the same you can be sure of the accuracy. If they don't then an average can be taken making the result more satisfactory.* (SA81a)

> *The students need to get an idea of by how much, if at all, the measurements for d varies. If they discover that d does not vary significantly then they can take an average measurement over just two rolls. They definitely do need an average value, though, in order to increase accuracy and reduce the affects of their experimental errors.* (SA82a)

*Choice A: Undetermined paradigm*

The remaining codes, for choice A, were assigned when it was not possible to determine whether the point or set paradigm underlie the responses. These responses did not suggest what to do once all the readings were taken. The first grouping, coded UA40 to UA45, links several repeats to either checking or obtaining better measurements. Two sample responses are presented below:

> *If we want to be exact in our measurement, the reading should be taken as many times as there is time for, because this eliminates random errors.* (UA43)

> *It is better to repeat an experiment more than once to be certain that your readings are accurate. It is possible that unexpected conditions such as the wind could alter your reading so it is better to double or even triple check your results by repeating the experiments.* (UA41a)

Responses that used the phrases 'the answer' or 'the result' to describe the outcome of the measurements were coded UA60 to UA65. Since these phrases could be used equally to describe a mean value or a recurring reading, the paradigm used is unclear. An example is shown below:

*The more times an experiment is repeated, the more of a general idea you'll get for the result. A one-off experimental may have an error and then the error is part at the result.* (UA62)

*Choice B: Point paradigm*

Students who side with B do not think that any repetition is required. Students could argue successfully that several repeats are unnecessary if they have specific knowledge that identical results would be obtained. At entry to university, students would most probably have no prior experience of the posited experimental setting. They would thus have no prior knowledge to suggest that one measurement is adequate. This procedural decision is thus considered to arise from the point paradigm. Consequently, responses that indicated agreement with cartoon character B were categorised as point. Only one response from the cohort in this study was coded PB01 (written text not codeable), the rest coded PB30 or PB31. Some examples of the latter are quoted below:

*Rolling the ball a few more times won't change anything, the height (h) is not changed.* (PB30)

*Assuming ideal situations (which we do a lot of in physics) e.g. no wind and homogenous surfaces, the ball would always land in the same spot.* (PB31)

Code PB40 would be assigned for answers that acknowledged the possibility of obtaining different results, but judged the exercise of repeating to be either confusing or pointless. These students clearly did not know how to deal with the spread in data sets. Some students enter university with strongly held perceptions. Responses that stress that repetition is a waste of time or resources would have been coded PB50. There were no responses from this cohort that were coded PB40 or PB50.

*Choice C: Point paradigm*

Respondents who indicate agreement with C reason that the best procedural action is to repeat the measurement once only. Typically, responses mentioned the need to verify or confirm the first reading. The quoted responses below were thus coded PC30 to PC35, following the same assignment methodology described before:

*Even though I think that B is right in that we don't need to do any more rolls from the same height, I feel compelled to do it just one more time to check the standard that we are going to be using i.e. h = 400 mm. Rather safe than sorry.* (PC30)

*Measurements can often be inaccurate due to bad methods of reading the measurement. The distance d should not change no matter how many times the ball is rolled down. Doing it at least one more time is to make sure that the original measurement was right.* (PC32)

*There may be slight inaccuracies within the first measurement of d = 436 mm, or at the point of release of the ball. Therefore, a second measurement has to be taken to make sure d is correct. However a third or fourth attempt is usually not necessary if one is precise in one's experiments.* (PC35p)

Responses that indicate additional readings would be required if the second result differed from the first were coded PC38 as in the sample response below:

*One must confirm that the measurement d is correct by rolling the ball again as, an error could have accidentally be made. Only if the second answer does not equal the first should it be rolled again, otherwise one would be wasting time.* (PC38et)

PC50 is similar to PB50, whilst PC60 would have been assigned for students who hope to identify the correct measurement from only two readings, with no indication as to how this choice would be made. No student in this cohort gave such responses.

*Choice C: Set paradigm*

Responses were coded SC81 where it was proposed that averaging of the first two results is sufficient if the numerical difference is not too great, and more readings required only in the case of significant deviation:

*The students should roll the ball one more time to check that no error was made the first time. If the two answers are fairly similar, an average could be taken. If the second answer is totally different the ball should then be rolled a third time.* (SC81)

Responses that mentioned the need to take more readings and calculate an average only if the second result differed from the first were coded SC82 as below:

*Well, the surface of the slope is likely to have various irregularities, and other variables may come into play, which together would affect the result i.e. results may not be reliable, and may differ should the same experiment be carried out once again. Thus, to make sure there aren't any variables wildly affecting the results, the experiment should be repeated as a control or confirmation of the result.*

*However, should this second result differ significantly, then A would be right, and an average should be taken from at least 3 measurements.* (SC82x)

*Choice C: Undetermined paradigm*

It is possible that some responses that side with C cannot be classified within the point or set framework. Typically, these responses do not indicate whether a recurring measurement or average is sought. The codes UC40-43 concern statements about the results, whereas UC80 was assigned when a decision was made based on the relative deviation of the second result. There were only two examples in this cohort:

*There is not much need to repeat the process over and over again as there is not much of errors which can be made by experimenters like say when measuring time with a stopwatch. There is a possibility of having areas in stopping and starting a clock which then requires several measurements to take the average. The one more time repeat only, is simply making sure of the result roughly because if one gets a very big difference in distance, there certainly would have been an error somewhere.* (UC42s)

*Choice C allows the students to check that the result they obtained with the first roll is reasonable. If the values differ by a large amount, then the first and second results should be repeated / redone.* (UC80)

Table 3.1 shows the frequency of categories used to code student responses to the RD probe from this cohort, before and after instruction. Related categories are grouped, e.g. all responses categorised SA20-26 were grouped as all suggest that repeated measurements are needed to calculate a mean. The coding scheme described in detail above was developed for all students from all cohorts. The table (and subsequent ones for the rest of the probes) thus includes only code groups that were used for the students involved in this study, i.e. mainstream students that completed both questionnaires prior to and after instruction. Table 3.1 summarises how the mainstream students acted and reasoned when making procedural decisions for distance measurements.

*Table 3.0.1: Summary and frequency of code categories established for mainstream students' responses to the RD probe before and after instruction. (n =53)*

| Codes | Descriptor | Pre-course | Post-course |
|---|---|---|---|
| PA30 | Several repeats to confirm a recurring reading. | 0 | 1 |
| PB01 | No repeats, no written response. | 1 | 0 |
| PB30-31 | No repeats, one single measurement adequate. | 4 | 0 |
| PC30-38 | One repeat only to confirm first result. | 7 | 0 |
| SA20-26 | Several repeats to calculate a mean value. | 33 | 38 |
| SA70-75 | Several repeats to gauge the uncertainty in the measurement result. | 0 | 3 |
| SA80-82 | Several repeats to confirm a recurring reading or calculate mean if spread occurs. | 1 | 0 |
| SC81-82 | One repeat and calculate mean from two results; if significant deviation in 2nd result, repeat several more times. | 2 | 0 |
| UA40-45 | Several repeats to check reliability of readings. | 3 | 6 |
| UA60-65 | Several repeats to check reliability of final result. | 2 | 5 |
| | **Total** | **53** | **53** |

### 3.3.2 Repeating distance measurement again probe (RDA)

Probe RDA solicits responses from students concerning the purposes of repeating measurements (distance) in the instance that a different reading from the first (result of 2nd roll) is observed. The text of the probe sheets follows:

The group of students decide to release the ball again from $h = 400$ mm. This time they measure $d = 426$ mm.

First release: $h = 400$ mm $d = 436$ mm

Second release: $h = 400$ mm $d = 426$ mm

The following discussion then takes place between the students:

**A**: We know enough. We don't need to repeat the measurement again.

**B**: We need to release the ball just one more time.

**C**: Three releases will not be enough. We should release the ball several more times.

The coding scheme is presented below:

*UN00 - No response*
*UU00 - Uncodeable*

***I agree with A because ...***

*PA01* - *Uncodeable reason*
*UA20* - *Take the average of the two results*
*UA23* - *The result can be estimated by averaging two results*
*UA25* - *Average of two results yields reasonably accurate result*
*PA30* - *Result (d) unchanged for same height [should not change]*
*PA40* - *Repeating will give a different result again, no point in repeating*
*PA41* - *Difference due to external factors*
*PA45* - *Accuracy is acceptable (results valid)*
*PA50* - *Repeating is a waste of time/resources*
*PA60* - *No point in repeating - h is independent of d*
*PA61* - *No point in repeating as we know that h is independent of d since all ext conditions constant*
*UA75* - *Result can be estimated in the range*

***I agree with B because ...***

*UB01* - *Uncodeable reason*
*PB10* - *You need to practice.*
*PB15* - *Practice will get a more accurate/better measurement*
*SB20* - *3 are needed to take an average/mean (answer/result better)*
*SB21* - *Get average/mean to reduce effect of outside/random factors*
*SB22* - *Get average/mean to reduce effect of "errors"/mistakes*
*SB23* - *Get average/mean and a (better/narrower) spread/uncert/std dev*
*SB24* - *To get an average to get closer to the true value*
*SB25* - *3 suffice for reasonable acc average*
*SB28* - *3 measurements to get average; more measurements needed depending on margin of diff*
*PB30* - *Get the recurring/correct answer - be confident of answer*
*PB31* - *Confirm recurring/correct reading – determine reason (ext factor) for difference*
*PB32* - *Confirm recurring/correct reading - eliminate "errors"/mistakes*
*PB35* - *Confirm recurring/correct reading - check accuracy (inaccuracies in measuring)*
*UB40* - *3 is enough/needed for validity (surety); too many different readings confusing*
*UB42* - *Ensure deviation not due to expt error (-if 3rd result similar, validity is confirmed)*
*UB45* - *Confirm the accuracy of measurements (two estimates are not enough)*
*PB50* - *Many repeats is a waste of time/resources*
*UB60* - *Get a general/reasonable solution/result/measurement*
*UB62* - *Get a more /reliable result-reduce effect of "errors"/reduce margin of error*
*UB63* - *3 is sufficient for unbiased result as probability of large dev in many repeats small (due to time)*
*UB65* - *To be more accurate/get a more accurate answer*
*SB70* - *To gauge the variance in the data (how 3rd differs from first two)*
*UB80* - *Decide to do more measurements if variance of third result is great cf first two*
*SB81* - *Decide to take recurring value or an average if variation*
*SB82* - *Indicate incorrect measurement, or calculate average if 3rd close to first two. If deviation large, many more measurements needed for average.*

***I agree with C because ...***

*UC01* - *Uncodeable reason*
*PC10* - *You need to practice.*
*PC12* - *Several releases will reduce the inconsistency in measurements*
*PC15* - *Practice will get a more accurate/better measurement*
*SC20* - *More measurements are needed to take an average/mean (answer/result better)*
*SC21* - *Get average/mean to reduce effect of outside/random factors*
*SC22* - *Get average/mean to reduce effect of "errors"/mistakes*
*SC23* - *Get average/mean and a (better/narrower) spread/uncert/std dev*
*SC24* - *Get average/mean to get closer to the true value*
*SC25* - *Get an average to be more accurate/get a more accurate answer*
*PC30* - *Get the recurring/same/correct answer - be confident of answer*
*PC31* - *Confirm recurring reading - eliminate outside factors that caused deviation*
*PC32* - *Confirm recurring reading - eliminate "errors"/mistakes*
*PC35* - *Confirm recurring reading - check accuracy*
*UC40* - *More repeats needed to confirm/be certain of measurements*

*UC41* - *Investigate and study nature of factors that causes the deviation in results*
*UC42* - *Take account of /investigate/minimize errors/inconsistency that caused deviation*
*UC45* - *Confirm accuracy of results /obtain accurate measurements*
*UC60* - *More measurements required to obtain acceptable/correct/conclusive result*
*UC61* - *Eliminate/take into account outside factors on result*
*UC62* - *Reduce effect of "errors" on result (to get better/more reliable result)*
*UC63* - *Determine if any of the meas'ts were incorrect and/or formulate the "most probable" distance of the readings*
*UC64* - *Get closer to the true value*
*UC65* - *To get a more accurate/(statistically valid) answer/result*
*SC70* - *To gauge/determine the spread/uncertainty*
*SC71* - *To gauge/determine the spread/uncertainty and investigate reasons (ext factors) for deviation*
*SC72* - *To gauge spread + more repeats might yield a smaller range/identify anomalies (due to errors)*
*UC80* - *Decide to take recurring reading or take more measurements if variation*
*SC81* - *Decide to take recurring (/accurate) reading or an average if variation (d=investigate reasons for dev)*
*UC95* - *More accurate scientific theories and conclusions can be drawn*

The response codes for siding with cartoon character A in the RDA probe are similar to those indicating agreement with B for the RD probe. Students that choose A do not see the value of repeating measurements, even in the light of a different reading obtained for a second release of the ball.

*Choice A: Point paradigm*

A student might refuse to acknowledge the different result. Responses that make a statement of a perceived physics fact, that *d* should be unchanged for the same height *h*, as reason for not performing more repeats should be classified within the point paradigm, coded PA30. There were no examples from this cohort. Similarly, the deviation in the second result may have been viewed as a refutation of the aforementioned "physics fact or theory" and hence any further repeats were deemed meaningless, as in code categories PA60-61. One student gave the following response:

> *It's useless to re-drop the ball from the same height if we already know that d is independent of h because we dropped the ball under exactly the same conditions of speed (since we didn't push the ball) and friction (we used same room).* (PA61)

Statements made about the deviation of the two results whilst agreeing with A, were coded within the point paradigm, as in codes PA40-45. Below is a sample response:

> *There was a difference of only 10 mm between the two answers. Due to the nature of the experiment, it would be unreasonable to expect better accuracy. A deviation of only about 2% should be small enough for the results to be considered valid.* (PA45d)

Response category PA50 is similar to PB50 in the RD coding scheme. There are no sample responses for this category.

*Choice A: Undetermined paradigm*

Students might believe that two readings are sufficient to calculate a mean value. These students have no way of gauging the spread in readings from multiple repeats, hence cannot argue convincingly that the deviation observed is "small" or "acceptable" nor can they defend their conviction that a mean calculated in this way may be a reasonable reflection of a measurement result. On the other hand, electing to calculate a mean is a set action. Therefore, the paradigm underlying these responses is not clear, point reason and set action, hence a U (undetermined) paradigm code would have been assigned as in UA20-25. No responses for this cohort were so classified.

UA75 would have been assigned if a response suggested that a result might be estimated within the range of the two readings. The reasoning could be considered as consistent with the set paradigm (an appreciation for spread in data sets) but believing this is possible by taking only two readings is a point action (two data points enough to determine the boundaries of a range). Again, there are no sample responses.

Responses for agreeing with cartoon character B in the RDA probe are similar to those given when siding with C in answering the RD probe. Students who make this procedural decision do not view it necessary to repeat the measurement beyond one additional release of the ball.

*Choice B: Point paradigm*

The "practice" codes PB10-15 would have been assigned for responses similar to those for RD described in section 3.3.1. No responses from this group fell in this category.

Responses that are coded PB30-35 indicate the need for a third release to identify the correct reading from the first two. Again, the assignment follows the pattern as described earlier. The response below was typical:

> *To release the ball a third time it would help to see which of the measurement is out when the ball is released for the third time.* (PB30)

Code PB50 would have been assigned for answers similar to that coded PB50 and PC50 in the RD probe. The current group of students did not write such responses.

*Choice B: Set paradigm*

Student responses indicating that a mean should be calculated from three releases (choice B) were coded SB20-28 following the same code assignment as described in section 3.3.1. The choice to repeat once only may be viewed as problematic since there is no clear indication that three readings would adequately account for the spread in the data set. However, these responses are believed to be consistent with the set paradigm since students often enter university having learned procedural rules of thumb at school, as evidenced in the sample responses below:

> *Three readings to obtain an average is enough.* (SB20)

> *Repeating the experiment 3 times should get an accurate answer. It is not practical, time wise to repeat the experiment too many times – it would take too long and there should not be too great a difference in the average of the first three results and the average of say, six results.* (SB25t)

Some responses justified the procedural choice of three readings by arguing that the deviation in the first two releases is small (SB20d), while others left the door open for additional repeats based on the outcome of the third try (SB28):

> *The measurements of d in both the first and second releases were quite close, so I think that a third (and final) measurement should give a good average d.* (SB20d)

> *Yes a difference might have been noticed and so the need for three releases will be enough to average the results. It will all entirely depend on the margin of the difference in the results, whether more values need to be measured or not.* (SB28d)

The reasoning in SB70 is similar to that for SA70 in the RD probe. Below is a sample response:

> *To see how different the third value will be from the two that they have.* (SB70)

Codes SB81-82 were assigned for responses to the RDA probe for similar reasoning as that for SC81-82 in the RD probe; an example is shown below:

*It is possible that one of the measurements was misread off the metre stick (the fact that they differ by 10 mm / 1 cm – these things happen!!). A third measurement may indicate such a mistake, or should fall close to these two measurements. Then an average can be taken. However, should they differ greatly, C would be right, as the more measurements taken, the more accurate the average would be.* (SB82ed)

*Choice B: Undetermined paradigm*

The codes UB40-45, UB60-65 and UB80 were assigned for answers similar to those described in section 3.3.1 where no procedural action was suggested (identical numerical codes). The written response below is representative:

*More tests would be good for coming to an unbiased result, but considering the time one has and the small chance that more tests could have dramatic difference, 3 readings should be sufficient.* (UB63t)

The procedural decision to take many repeats, choice C, is identical to choice A in the RD probe. The codes for choice C are thus nearly identical to those for choice A in the RD probe, with a few additions to accommodate the range of responses. Many responses given by this group of students mentioned the need to repeat many times to deal with the large variation observed in the second release.

*Choice C: Point paradigm*

No examples for codes PC10-15 exist for this cohort. Even though a sizable deviation is evident in the second reading, students might still believe that it is possible to extract one recurring or correct measurement (PC30-35). The responses below typify this thinking:

*One needs to find out which value is correct. By rolling just one more time, the same error as before could be made again therefore rolling several times will hopefully confirm the correct value.* (PC32)

*Because of the difference between the first and second results it is necessary that the ball should be released several more times until such time that same results are obtained.* (PC30d)

*Choice C: Set paradigm*

The reasons given and therefore the assigned codes in the sample responses below are very similar to those given in the RD probe:

*It is now apparent that the experiment does not give the same result every time, and so multiple experiments should be done, and the average of the results should be used.* (SC20d)

*d – deduced in first release is 436 and this has a 10 mm variance value to that of 2nd release and this may either mean that 1st / 2nd reading has an error. Therefore taking more readings will help us know which one of the two is more reliable and we can thus calculate average of numbers that are closely related and constitute majority part of readings taken.* (SC22d)

*There is quite a big difference between the two measurements (1cm). To have a truly accurate average, repeating the experiment several more times would give one a better idea of the differences between various results.* (SC25)

*At least three measurements are required in order to gauge the variance of d. Only C provides for this. Already a 10mm difference has been established between the first and second measurements: there is a significant variance.* (SC70d)

*The results are quite different, thus for the sake of accuracy more measurements should be taken. If only one more was taken it might not help, if the result was very different. Several more should be taken to work out a convincing average or to identify one repeating result.* (SC81ad)

*Choice C: Undetermined paradigm*

The code categories UC40-45, UC60-65 and UC80 are similar to the corresponding categories described previously for the RD probe. Code UC95 was assigned when a response proposed that more "accurate scientific theories" or better conclusions result from more readings. Since no indication is given how to deal with the spread, the underlying paradigm is not obvious. One student gave the following written text:

*The more information can be recorded, the more accurate the scientific theories and conclusions can be drawn.* (UC95)

Table 3.2 presents the code categories used by the mainstream students both before and after their laboratory course when dealing with repeated distance measurements.

*Table 3.2: Summary and frequency of code categories established for mainstream students' responses to the RDA probe before and after instruction. (n =53)*

| Codes | Descriptor | Pre-course | Post-course |
|---|---|---|---|
| PA40-45/ PA60-61 | No additional repeats necessary based on deviation observed. | 2 | 0 |
| PB30-35 | One additional repeat to confirm correct reading. | 4 | 0 |
| PC30-35 | Several repeats to be confident of a recurring reading. | 1 | 1 |
| SB20-28 | One additional repeat to calculate a mean value. | 5 | 4 |
| SB70 | One additional repeat to gauge the spread in the data. | 0 | 1 |
| SB81-82 | Confirm correct reading or calculate a mean based on variance of 3rd reading; more repeats indicated if deviation large. | 1 | 0 |
| SC20-25 | Several repeats required to calculate a mean value. | 22 | 23 |
| SC70-72 | Several repeats to gauge spread/ uncertainty in data. | 1 | 4 |
| SC81 | Take recurring reading or mean value if variation in readings. | 3 | 0 |
| UB40-45/ 60-65 | 3 reading sufficient to be confident of readings/ "result". | 1 | 3 |
| UB80 | More readings if variation in 3rd compared to first two is large. | 1 | 1 |
| UC40-45/ 60-65 | Repetition required for confidence in readings/ "result". | 11 | 15 |
| UC01 | Several repeats, no written justification given. | 1 | 1 |
| | **Total** | **53** | **53** |

### 3.3.3 Repeating time measurement probe (RT)

The structure of the RT probe is identical to the RDA probe. The text below summarises the contents of the probe:

> The students are now given a stopwatch and are asked to measure the time that the ball takes from the edge of the table to hitting the ground after being released at $h = 400$ mm. They discuss what to do:
>
> **A**: We can roll the ball once from $h = 400$ mm and measure the time. Once is enough.
>
> **B**: Let's roll the ball twice from height $h = 400$ mm, and measure the time for each case.
>
> **C**: I think we should release the ball more than twice from $h = 400$ mm and measure the time in each case.

The cartoon characters' discussion comments are different but the same procedural decisions are presented as in RDA: A for no repeats, B for two repeats and C for several releases. The coding scheme is thus very similar to RD and RDA, hence only codes not explained before will be discussed here. The scheme is presented below:

*UN00* - *No response*
*UU00* - *Uncodeable*

***I agree with A because ...***

*PA00* - *No reason*
*PA01* - *Uncodeable*
*PA30* - *Repeats will give the same result (time depends on constant variables)*
*PA31* - *Repeats will give the same result if outside factors are constant*
*PA35* - *Repeats will give the same result if a very accurate measuring system is used*
*PA40* - *Repeats will give different results which is confusing/pointless*
*PA50* - *Repeats are a waste of time/resources*

***I agree with B because ...***

*UB00* - *No reason*
*UB01* - *Uncodeable*
*PB30* - *Confirm the (recurring/same) measurement*
*PB32* - *Check that no "errors" was made in obtaining (recurring/same) measurement*
*PB35* - *Check the accuracy of the (recurring/same) measurement*
*PB38* - *Check the validity of the answer – if different repeat more until get same answer*
*UB40* - *Need a variety of/different results*
*PB50* - *Many repeats are a waste of time/resources*

***I agree with C because ...***

*UC00* - *No reason*
*UC01* - *Uncodeable reason*
*PC10* - *You need to practice.*
*PC11* - *Practice to minimize/take into account outside factors*
*PC12* - *Practice to eliminate "errors"/mistakes/discrepancies*
*PC15* - *Practice will get a more accurate/better measurement*
*SC20* - *Get average/mean (answer/result better/more valid)*
*SC21* - *Get average/mean to reduce effect of outside/random factors*
*SC22* - *Get average/mean to reduce effect of "errors"/mistakes*
*SC23* - *Get average/mean and a (better/narrower) spread/uncert/std dev*
*SC24* - *To get an average to get closer to the true value*
*SC25* - *Get an average to be more accurate/get a more accurate answer*
*SC28* - *Eliminate/discard erroneous results - average of results in similar range*
*PC30* - *Repeat until you get a recurring/same/correct answer - be confident of answer*
*PC32* - *Confirm recurring reading - eliminate "errors"/mistakes*
*PC35* - *Confirm recurring reading - check accuracy*
*UC40* - *Need many/different results / to be confident of results*
*UC41* - *Rule out/take into account outside/random factors*
*UC42* - *(Need at least 3 results) to make adjustment (discard anomalies) for expt error*
*UC45* - *Repeats are needed to confirm accuracy / be more accurate*
*UC60* - *Repeat to get a better answer/result*
*UC61* - *Get a more accurate/reliable result-take into account outside factors*
*UC62* - *Get a more accurate/reliable result-reduce effect of "errors"*
*UC63* - *Repeats are needed to get an approximate answer*
*UC64* - *Get closer to the true value*
*UC65* - *To get more accurate answer/result (easier to confirm)*
*SC70* - *Gauge the variance/spread of time*
*UC80* - *Decide to do more measurements depending on variation*
*SC81* - *Decide to take an average because of variation*

Previously it was explained that the RT probe, used only in the questionnaire before instruction, tests the role of context on procedural issues. The codes discussed and sample responses quoted

thus highlight the differences in reasoning used by students when dealing with time and distance measurements.

Students might believe that time measurements are not subject to deviation. Prior to answering the RT probe, they would have been confronted by deviation in the distance measurements presented in preceding probes (RDA, UR, AN, SMDS and DMSS). These students argued either for no repeats or one additional release to confirm the first reading as the following responses showed:

> *They should do it only once. Time will not be affected by how far the ball travels after leaving the board. Distance will depend on speed of the ball and acceleration while time will remain the same.* (PA30)

> *Whereas the distance travelled by the ball might be affected by external conditions, the time taken is unlikely to change. Time should remain constant. Therefore 2 attempts are enough. The second just to ensure that the first attempt was legitimate.* (PB30)

Students would have seen probes that explicitly deal with the determination of a mean before answering the RT probe. Responses could thus explicitly refer to the deviation seen for distance measurements in arguing for several time measurements as in the response below:

> *After the first question, I realised that each reading can be different. Thus, doing an experiment more than twice should give the best answer.* (UC60d)

Some students mentioned factors, such as the reaction time related to operation of the stopwatch as justification for taking many time readings. The code qualifier "s" was appended to the codes for such responses, as was the case for the following:

> *It is pretty difficult to know at exactly which moment to press a stopwatch button. If a different person timed each instance of the experiment, you are likely to get very different results. Even the same person could react a little more slowly in one instance than in another.* (UC41s)

> *We've seen from the previous experiments that there is a margin for error in the experiment, and it should be tested a few times. There is also more margin for error in the timing process. It relies on human hand-eye co-ordination, and so an average should be taken in any case.* (SC22sd)

The code categories identified for this cohort when answering the RT probe prior to any instruction are presented in Table 3.3.

*Table 3.3: Summary and frequency of code categories established for mainstream students' responses to the RT probe before instruction. (n =53)*

| Codes | Descriptor | Pre-course |
|---|---|---|
| PA30-35 | No repeats necessary – the same result will be achieved. | 2 |
| PB30-35 | One additional repeat required to confirm 1st reading. | 1 |
| SC20-28 | Several repeats to calculate a mean value. | 28 |
| SC70 | Several repeats to gauge the spread/ uncertainty in the time measurements. | 1 |
| UC40-45 | Many repeats required to be confident of readings. | 12 |
| UC60-65 | Multiple repeats required to obtain a reliable "result" | 7 |
| UC00 | No response (indication that students "ran out of time") | 2 |
| | **Total** | **53** |

### 3.3.4 Using repeated distance measurements probe (UR)

This probe does not provide any procedural options. It is open-ended in terms of respondents' decisions about dealing with and processing readings in a data set. Respondents have to make a procedural decision by providing a single number to represent the data set, which for the UR probe consists of distance readings, $d$ obtained by releasing the ball five times from the same height, $h$. The essential text of the probe is presented below:

> The students continue to release the ball down the slope at a height $h$ = 400 mm. Their results after five releases are:
>
> | Release: | 1 | 2 | 3 | 4 | 5 | |
> |---|---|---|---|---|---|---|
> | Distance (mm): | 436 | 425 | 440 | 425 | 434 | |
> | | | [426] | [438] | [426] | | [post questionnaire values] |
>
> The students then discuss what to write down for $d$ as their final result … student discussion follows
> Write down what you think the students should record as their final result for $d$.

The coding scheme is thus constructed around the procedural decisions made and is presented below.

***A: mean = $\Sigma x_i / n$ ( i = 1 to 5 ) (432 mm)***

*SA20 - Average/mean of all the readings best*
*SA21 - Average/mean takes into account outside/external factors*
*SA22 - Average/mean takes into account errors/mistakes / provided no ext forces interfere*
*SA23 - Average/mean +/- standard deviation or uncertainty best represents data*
*SA24 - Average/mean is closer to true/expected/actual value*
*SA25 - Average/mean is more accurate*
*SA26 - Average/mean is closest to all values*
*SA27 - Average/mean + more repeats to check if outliers are valid*
*SA28 - Average best final result as it would take many repeats to find d that occurs most*
*SA30 - No clear recurring measurement (no result appears every time) → calc. mean*
*SA31 - Not possible to be sure which value for d is correct → average/mean best*
*SA32 - Results do not gravitate to particular value → equal error in each d → calc. mean*
*SA60 - Better to use average; takes into account all measurements, all valid /less biased*
*SA61 - No reason to ignore any reading/ all equally valid → average includes all*
*SA62 - No large "errors"/"duds- cannot discard any measurement → mean includes all*
*SA63 - Deviation due to random factors→ can't favour/discard any→ mean+/- std dev*
*SA64 - No experimental result is exact (=true value) → use all data to obtain average*
*SA70 - Average best represents true value as most measurements would fall within range (spread/uncertainty) of the mean*
*SA71 - Average is a good approximation/prediction of subsequent measurements*
*SA80 - Variation in measurements → average is a better solution/answer*
*SA81 - Variation in measurements small (no large discrepancies) → safe to take average*
*SA83 - Variation not large → differ due to random variation → average/mean*
*SA85 - Variation in measurements → average is most accurate value*

***B: mean = $\Sigma x_i / n$ ( i = 1 to 4 ) (430 mm)***

*SB20 - Disregard most wayward measurement*
*SB21 - Disregard most wayward, which is probably due to external factors*
*SB22 - Discard most wayward (furthest from average) measurement due to error/mistake*

***C: mean = $\Sigma x_i / n$ ( i = 1 to 3 ) (431.7 mm)***

*SC20 - Discard highest and lowest measurements*
*SC22 - Discard highest and lowest measurements as due to experimental error/mistakes*

***D: Recurring (425 mm / 426 mm in post)***

*PD30 - Take the measurement that recurs most often / appears more than once*
*PD81 - The 2nd & 4th releases agree OR find mean/average of all the readings*

***E: Best (434 mm)***

*PE34 - Take measurement that is closest to the average (432 mm)*
*PE40 - Take the last measurement – (deviation due to mistakes/external forces/factors)*

***F: Estimation / Range***

*SF70 - Range within which measurements are likely to fall as d different each time*
*SF71 - Approximate value of d in range of measurements ( e.g. d ≅ 430 mm )*

***U: Uncodeable***

*UU00 - Uncodeable answer*

Investigating responses to the UR probe from diverse groups of students yielded six different procedural actions to present a final result from a data set of varying readings. One of the letters, A to F is assigned according to which of these options is chosen.

*Action A: Set paradigm*

The letter A was assigned when a student opted to calculate the arithmetic mean of the five given readings. This action is indicative of the set paradigm; hence S was prefixed to the response codes. There are a few main arguments that could be used to justify this course of action, the numeric part of the code representing similar reasoning types as described for the data collection probes. The codes SA20-28 were assigned to responses that argued that a calculated mean value was the best way to record a final result. Various experimental issues were used in the arguments; external factors (SA21), 'errors' (SA22) or accuracy (SA25). Respondents could have argued that the mean is closest to a true value (SA24) or closer to all the values (SA26). Others might have argued that additional repeats were required to test the validity of outlying readings (SA27) or to determine which reading appeared most frequently (SA28). Actual written responses with the assigned codes are presented below as examples of the range of responses given by this cohort of students:

*I decided to take the average of all the results as my final answer.* (SA20)

*I would add all 5 measurements and then divide it by the number of attempts (5) to average the result. I would do this to try and reduce the effect of those unaccounted for variables that are influencing the experiment over which I do not have control.* (SA21)

*We can get quite an accurate result by adding the 5 results together and dividing by 5 to find the average (mean). That way, if there was experimental error in one of the results, it would get divided by 5 and not affect the final results greatly.* (SA22a)

*For the best possible answer the results should be worked out by finding the average and also recording the standard deviation for later computations. This answer is best because it gives the best representation the data, and it is more accurate than merely taking the average. Though at our level such accuracy is hardly ever required so the average would also be acceptable.* (SA23a)

*The average value is closest to the true value. This holds true for several readings i.e. average of two readings not as accurate as average of 5.* (SA24)

*An average or the mean of the 5 measurements gives the most accurate final result.* (SA25)

*To take the average of all the numbers is better than taking one number. An average is the closest number to all the results, where one number could be very different from the other results.* (SA26)

*I think that they should record the average of the five readings as the value for d. However, in this case where 426 is quite different to 438, I would consider rolling the ball some more to see if perhaps some readings are very inaccurate.* (SA27)

*The answer above is an average of the 5 measurements obtained. Only by repeating the experiment is it possible to find a number for d that occurs the most. Still taking the average is the best course of action.* (SA28)

Some students argued that it is necessary to calculate a mean as no clear recurring reading is observed after several repeats. These were coded SA30-32 as those below:

*They did not get the small value (425 mm) for d every-time they released the ball.* (SA30)

*The average value of d is appropriate with some error bounds, as no one can be exactly sure of what d value is correct. The error bound should encompass all the individual values found for d.* (SA31)

*Because most of the results have not gravitated to a particular value, it is likely that equal error is present in each and thus they should each have equal importance-hence I would take the average.* (SA32)

Other students argued that all the individual measurements are equally valid for various reasons and hence all the readings should be used to calculate a mean value. The responses below were coded SA60-SA64:

*This calculated value of d is the average of the measurements, as this represents a value that is representative of all the readings. An average takes into account all readings, thus all data is incorporated.* (SA60)

*There is no reason why one should ignore any reading from the experiment, thus the average would be the best answer for d.* (SA61)

*The mean i.e.* $\frac{1}{n}\sum_{i=1}^{n} x_i$ *gives you the 'average' value for d. There don't seem to be any outliers to sway d's average value, so the mean should work well.* (SA62)

*These are random responses of the ball from that height. No specific values should be favoured but all of these readings tend to encompass an average/ mean. Thus, it is this value that would be most representative.* (SA63)

*The mean value assumes most realistically that no experimental result is exact (unlike the median or mode averages) and thus uses all the experimental data to obtain a result.* (SA64)

Some students thought that the mean is a good representation of the measurement result since the data from multiple repeats would be spread around this value. Such responses were coded SA70-71 as below:

*This is the average of the releases. Although this exact result has not been observed, it is likely that after a large number of trials we would see that most results fall within a certain range of this "true " value.* (SA70b)

*This is an average of the distances from the five different releases. This will be around about the distance it will be each time or it will be close to this distance.* (SA71)

Where the variation in the readings or judgements about the size of the variation were used as justification for thinking a mean value (for all the readings) is best as the final result of a measurement, the codes SA80-85 were assigned as for the responses below:

*The values are different, so it's best to take an average.* (SA80)

*None of the measurements are vastly different from the rest so the average should be taken from all 5 readings.* (SA81)

*None of the given values lie far away from the rest, so it appears they differ due to random variation. It makes sense to take an average.* (SA83)

*The measurements are different ∴ to obtain the most accurate measurement from all the information one must take an average of all the values.* (SA85)

*Choice B: Set paradigm*

An alternative strategy used by some students was to discard the reading that they considered to deviate most from the rest and calculated a mean from the remaining four measurements. Students' responses yielded two categories of responses, SB21 and SB22, with samples below:

*From two equal results at 426, we can see that the expected value is around 430, therefore we disregard the highest value which is probably a very inaccurate measurement (due to some unfavourable circumstances).* (SB21a)

*The most wayward measurement (440mm) should be disregarded. They should take the average of the remaining measurements.* (SB22)

*Choice C: Set paradigm*

Another course of action, consistent with the set paradigm, was to exclude the highest and lowest recorded readings in determining a mean value. This may be problematic since no clear idea of the spread in the data is possible with only five readings; however it demonstrates an appreciation that there is a normal distribution of data when recording multiple readings. The students gave the following responses:

*Remove the highest and lowest value and get the average.* (SC20)

*Discard the smallest and largest values as being the result of experimental error. Take the average of the remaining three [d= (436+425+434)/ 3= 431.6].* (SC22)

*Choice D: Point paradigm*

Respondents may believe that one individual measurement adequately represents a measurement result even with a clear spread in recorded data. The action to select the recurring measurement or the most repeated reading is rooted in the point paradigm (PD30). Certain students may not have been convinced with two identical readings after only five repeats and thus indicated the alternative to calculate a mean (PD80). Students from our sample gave the following explanations for their actions:

*They acquired this measurement for d more than once. They did not obtain the other values more than once. Thus, it is more than likely that 425 is correct. 425 is also a rough average of all the d measurements.* (PD30)

*The 2nd release and 4th release agree or we can find a mean value by adding all the distances and dividing by 5.* (PD80)

*Choice E: Point paradigm*

Another course of action is to select the reading (433 mm) that is closest to the arithmetic mean of all the recorded results (PE34). This reading happens to be the last recording in the list. Respondents may explain that measurement technique is perfected with multiple repeats and hence the last reading (434 mm in this case) be taken as the measurement result (PE40). Students here acknowledge spread but do not accept it as inherent in experimental results. They do not yet realise that a data set should be modelled (mean). Responses from this cohort are shown below:

> *It is the average mean value between 425 and 436.* (PE34)

> *As the students are performing the experiment, after each release they come to see that the previous one was wrong due to an external force on the ball, a wrong height, h, or a wind blow. The students come to get their lesson by their own mistake that is why they carry the experiment 5 times.* (PE40ex)

*Choice F: Set paradigm*

Some students did not provide a single number in the final result box on the answer sheet but indicated "a range" or an approximate number. These responses are deemed to stem from the set paradigm as no single recorded reading is chosen to represent the data set with an appreciation for the inherent uncertainty in a measurement result. The following responses were received:

> [Final result = *"A range"*] *The ball would probably travel a distance, which is in this range, each time it is dropped from 400mm. The range would be a good enough idea for how far it travels i.e. between 425 and 440, because each time the experiment is carried out, d will probably alter slightly.* (SF70)

> [Final result = *"$d \cong 430$ mm"*] *As we can see the value of d is between 425 and 440.* (SF71)

Table 3.4 shows the extent of the code categories identified for this cohort of mainstream students when dealing with a data set consisting of repeated measurements.

*Table 3.4: Summary and frequency of code categories established for mainstream students' responses to the UR probe before and after instruction. (n =53)*

| Codes | Descriptor | Pre-course | Post-course |
|---|---|---|---|
| PD30 | Choose the reading that appears most often. | 1 | 0 |
| PD81 | Choose recurring reading or calculate a mean value. | 1 | 0 |
| PE34 | Select the best reading closest to the mean. | 1 | 0 |
| SA20-28 | Mean value including all the readings. | 26 | 35 |
| SA30-32 | No clear recurring reading; therefore the mean is the best. | 3 | 1 |
| SA60-64 | Mean is best as it includes all the values, which are equally valid. | 7 | 9 |
| SA70-71 | Most values would fall within uncertainty range from the mean. | 5 | 0 |
| SA80-85 | Due to the variation, the mean is the best option. | 5 | 7 |
| SB20-22 | Mean calculated from four best reading; most wayward discarded. | 1 | 1 |
| SC20-22 | Highest and lowest readings disregarded in determining the mean. | 1 | 0 |
| SF70-71 | Final result is a "range" or an approximate number. | 1 | 0 |
| UU00 | No response ("ran out of time") | 1 | 0 |
| | **Total** | **53** | **53** |

### 3.3.5 Fitting a straight line graph probe (SLG)

Responses to this probe are coded according to how the trend in the plotted data points is modelled. Students are free to interpret the data points in any way they want. The codes for the different response categories were assigned based on firstly, the paradigm used, then the type of line drawn, followed by the fitting method used and lastly whether or not the line was forced through the origin. A letter code was thus assigned hierarchically for each of the procedural choices indicated in the responses. Image scans of students' actual drawn lines or curves on the given graph will be included with the written responses as the graph is critical in determining the response category.

The basic text of the SLG probe follows (graph of plotted points excluded):

> A group of students collect data at different heights and use it to plot a straight line graph. The data are plotted below. On this graph, draw the line that you think best fits this data. Explain carefully what you have done and why.

The coding scheme is presented below:

*Responses are coded with a combination of the following codes: Paradigm(P/S/U)-1-2-3*

***1. Type of line***
*C - Curve drawn*
*L - Straight line drawn*
*P - Points are joined*

***2. Fitting method***
*B - Line forced trough bottom point/points*
*F - Best fit (line that best fits all the points)*
*M - Line forced through middle points*
*T - Line forced trough top point/points*
*X - Line joins top and bottom points*

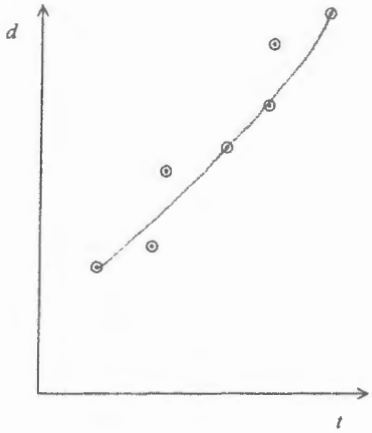
***3. Origin***
*O - Line forced through the origin*

The responses for this cohort showed no serious discrepancies between the graphs plotted and the reasoning that followed due to the high level of sophistication in written responses. It was therefore relatively simple to identify the underlying paradigm (point or set) used by the students with none coded U (undecided).

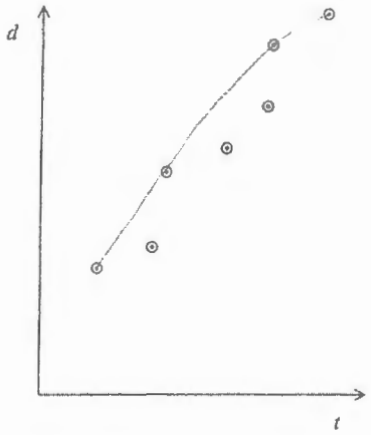
Actions associated with the point paradigm are those where either multiple line segments are forced through all the data points or a single line or curve connects selected data points. Respondents may have reasoned that particular data points best describe a desired trend (e.g. a straight line that joins the origin and data points which line up with the origin) or that each and every data point determines the trend by necessity.

*Point paradigm (P): Curve drawn (C)*

Respondents may have forced a curved line through data points that lie near the middle (PCM), top (PCX) or bottom (PCB) of the trend. The curve is a consequence of the particular data points chosen to represent the trend. For this group, only a few responses were coded PCM and PCX as those below:



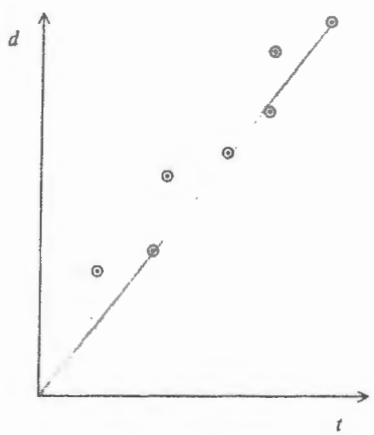

*This line seems to indicate the most accurate curve for the data as the most points fall on the curve. The points that are far from the curve could indicate inaccuracies or inconsistencies in the experiment, therefore it is not vital that they be on the graph.* (PCM)



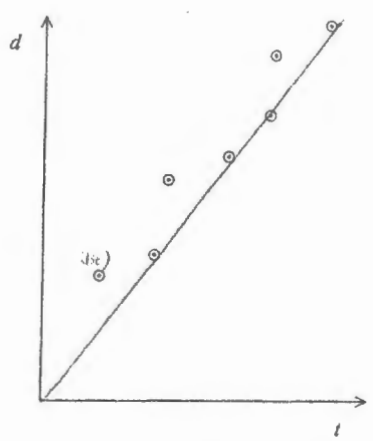

*What I have done is I have drawn a curve between the first point and the last point and the reason for doing what I have done is that when I check the results between the groups the points start with a small point and also ends with a small point - when represented graphically yields a curve.* (PCX)

*Point paradigm (P): Straight line drawn (L)*

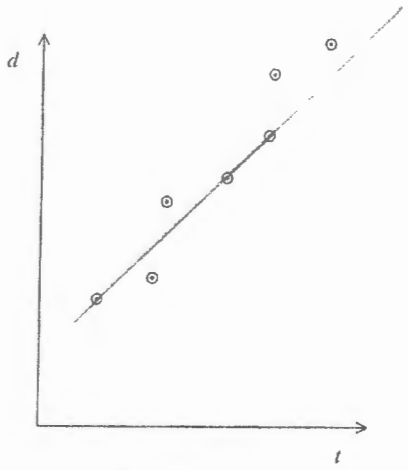
Some students forced a straight line through various combinations of data points as before (PLB, PLM, etc.). However, here the choice of data points might have been influenced by the inclusion of the origin (PLBO and PLMO). From these responses it was clear that it is imperative for the line to pass through the origin. For this cohort, response types PLBO, PLM and PLMO were identified as below:



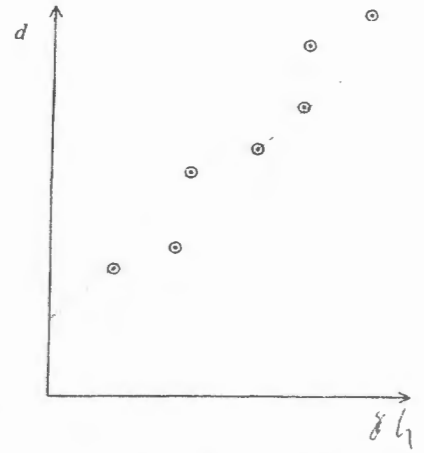

*A line is drawn in such a way that it includes as many of the data as possible to get the best possible idea of what the actual results should be.* (PLBO)



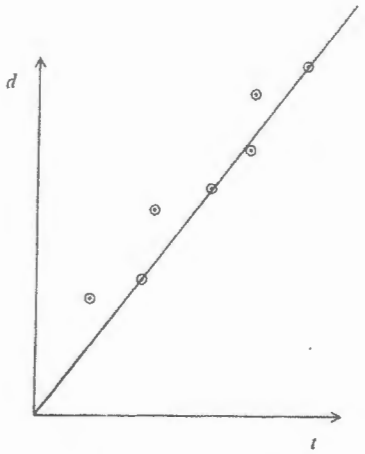

*This line corresponds with most of the values found. Also one can only cut the first dot if it is connected between the origin & itself.* (PLBO)



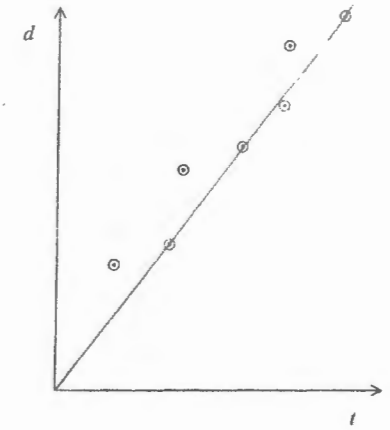

*The graph that should be obtained is a straight line and the best thing to do is to join as many points as possible and in this case the maximum I can join to fit a straight line is three which I have joined.* (PLM)




*They are the only group of at least three points in a straight line and the line has points on either side of it (it's in the middle).* (PLM)



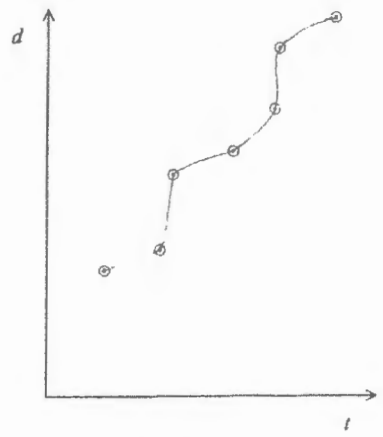

*I have drawn a line that starts at the origin which means at t = 0 the ball was at rest and the line also goes through the points that are close together which makes it more acurate and the line also shows that as the ball covers more distance, it takes more time.* (PLMO)



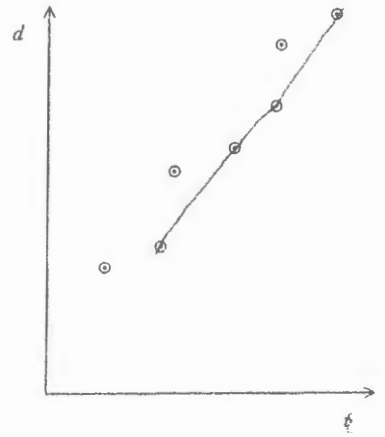

*This straight line graph represents the average of all their readings because this line passes through the most readings and therefore is the best average.* (PLMO)

*Point paradigm (P): Data points joined by line segments (P)*

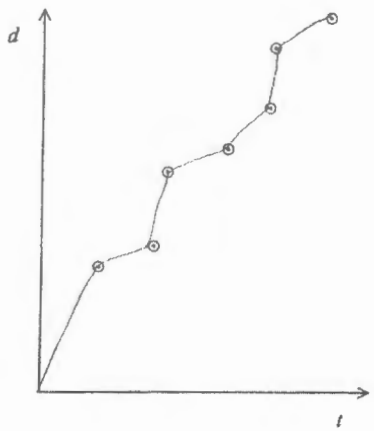
Instead of drawing one continuous line or curve, respondents here chose to join either all (PP) or selected (PPB, etc.) data points with line segments. A few responses received were coded within these categories:




*All points I have joined by a flexible line which is not necessarily straight because the change in time from point to point is not the same as the distances are also not the same. This is because of the variation in the intervals of distance and also of time.* (PP)




*Drawn d as a function of t to show the relationship of d to time.* (PPB)

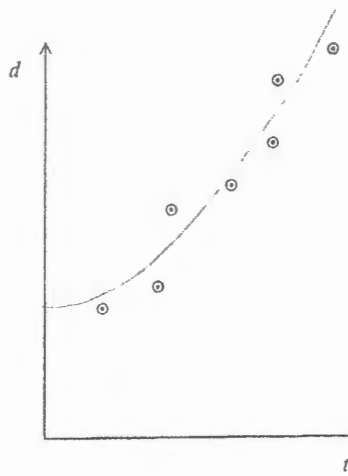



*The students aimed to plot a straight line graph, however some of their figures were inaccurate, hence the crooked line.* (PPO)

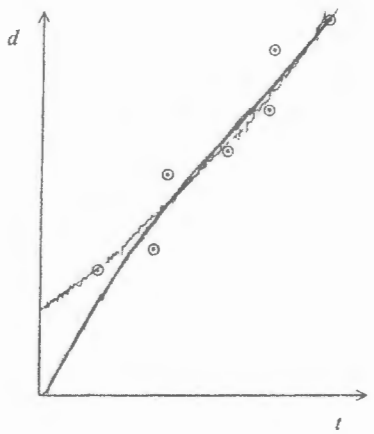
The procedural action of fitting a least squares straight line to all the data points is associated with the set paradigm. Respondents reasoned that modelling the trend in graphical data best represents the results of a series of measurements. Since the least squares analysis is not done explicitly, the lines drawn together with the reasoning determines whether or not a response is associated with the set paradigm.

*Set Paradigm (S): Curve drawn (C)*

Here, respondents opted for a curve instead of a straight line (SCF). This is possibly due to more plotted points being below the general straight-line trend. Including the origin may also influence the choice of a curve as some students might believe that the $d$ vs $t$ plot should yield a trend including (0,0). However, with these responses all the data points are taken into account in fitting the curve and are therefore coded as set. The following responses demonstrate this type of action:



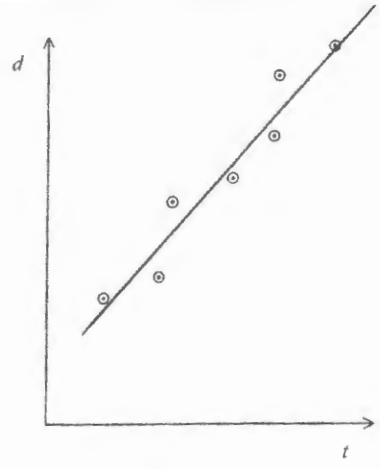

*The recordings do not show a continuously curving line of a parabola or any uniform motion. We must assume that the answers are slightly inaccurate but the general shape of the graph is accurate. Therefore a general line in the same area of the dots is drawn to represent this uniform motion. Practical experiments are not always entirely correct and require a certain amount of deviation.* (SCF)



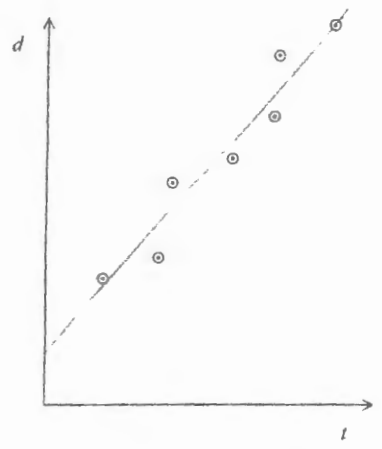

*The line has been drawn in such a manner that it encapsulates all the data and does not omit anything. The line has been drawn in relation to the general average result and thus does not pass through all the points. It is drawn in this way so that you can easily access the general average result in relation to the time.* (SCFO)

*Set paradigm (S): Straight line drawn (L)*

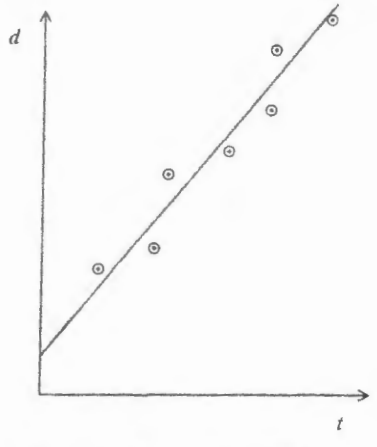
Responses where a least squares type fitting procedure was clearly applied to the data were coded SLF. Again, some students might have thought that $d$ is directly proportional to $t$ and hence drew the line through the origin, coded SLFO. In the versions of the questionnaire used in the pre- and post tests, the first and last data points fall close to the general trend line. It is then conceivable that some responses would indicate a fitting procedure but also connecting these points, coded SLFX. The written responses with accompanying graphs below is representative for this cohort:



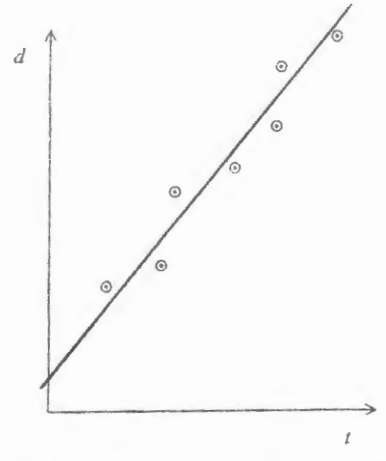

*The line drawn has been drawn in like an 'average' of all the points plotted, even though it does not go through any of the points, it is the line whose gradient will give the best representation of the results obtained.* (SLF)



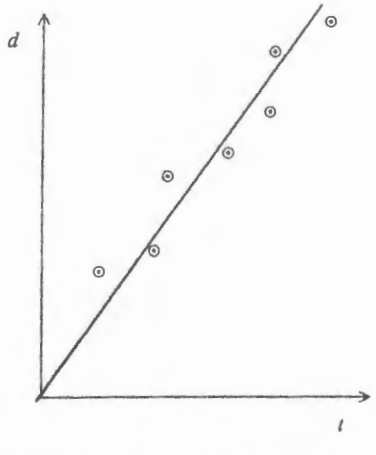

*I drew a straight line with the dots spread evenly on both sides.* (SLF)



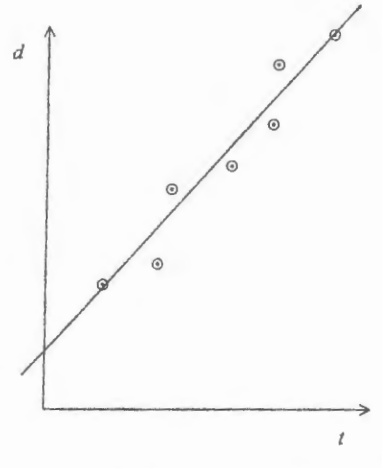

*The sum of the perpendicular distance deviations of the readings from the line are hopefully a minimum for this line. The line should not start at the origin because there is* $t > 0$ *for* $h = 0$. (SLF)




*The line was drawn using the principle of "Least squares", which states the sum of the squares of all the distances from each point to the line should be minimised. This is effectively equivalent to minimising the uncertainty in the result.* (SLF)




*The straight line would go in between the results plotted, representing the average. The results would not all be plotted on the line, but should be near to the line. The line would go through the origin since d is directly proportional to t.* (SLFO)




*I chose a line that would accommodate all points with a small uncertainty and by so doing I joined the first and last points.* (SLFX)

Table 1.3 in Chapter 1 presented the actions and reasoning associated with the point and set paradigms respectively in the various areas of measurement. The sample size of this cohort of students is relatively small. Consequently, responses were coded as set according to the underlying reasoning used, even though a point action was indicated when students forced their best-fit straight lines through certain points like the origin (codes SCFO, SLFO and SLFX). Figure 1.1 presented a schematic of the four areas within which a student may be located in terms of action and reasoning. Hence, even though the responses described here were coded as consistent with the set paradigm, the analyses in the following chapter will keep account of the fact that these students applied a point action in their responses.

Table 3.5 shows the range of response categories identified for mainstream students both before and after a laboratory course that included a data processing and data analysis component. The table summarises the actions and reasoning employed by these students when modelling a trend in plotted data points.

*Table 3.5: Summary and frequency of code categories established for mainstream students' responses to the SLG probe before and after instruction. (n =53)*

| Codes | Descriptor | Pre-course | Post-course |
|---|---|---|---|
| PCM | Curve forced through the middle data points: reasoning – procedure includes as many points as possible. | 1 | 0 |
| PCX | Curve forced through the first and last data points: reasoning – this will give the best representation of the trend. | 1 | 0 |
| PLBO | Straight-line forced through the bottom points: reasoning – these points line up with the origin. | 5 | 0 |
| PLM | Straight-line forced through the middle data points: reasoning – procedure includes as many points as possible | 2 | 4 |
| PLMO | Straight-line forced through the middle points: reasoning – these points line up with the origin. | 5 | 1 |
| PPB | Every data point joined by disjointed line segments: reasoning – all the data points must be accommodated. | 1 | 0 |
| SCF | Curve fitted: reasoning – it best represents the overall trend in the data. | 2 | 0 |
| SCFO | Curve fitted: reasoning – it best represents the overall trend in the data whilst including the origin. | 1 | 0 |
| SLF | Straight-line fitted: reasoning – it best represents the overall trend in the data. | 24 | 42 |
| SLFO | Straight-line fitted: reasoning – it best represents the overall trend in the data whilst including the origin. | 10 | 5 |
| SLFX | Straight-line fitted: reasoning – it best represents the overall trend in the data – coincidence that it passes through the 1st and last points. | 1 | 1 |
| | **Total** | **53** | **53** |

### 3.3.6 Comparing data sets with the same mean and different spread (SMDS)

The SMDS probe requires respondents to make a judgement on the relative quality of two data sets that have identical means but differing dispersion in the data. The point paradigm underlies responses where the choice of which group produced the best results depends only on whether the means are identical or similar. The spread in data is not seen as a result of the measurement and hence individual numbers, whether the mean or single readings, are compared with one another and judgements made based on similarity. The set paradigm is indicated where the choice of the better measurement is based on the relative sizes of the standard deviations (size of range or spread). The reasoning is that the spread is an indicator of the precision of the measurement. The text below summarises the contents of the probe sheet:

> Two groups of students compare their results for $d$ obtained by releasing the ball at $h$ = 400 mm. Their results for five releases are shown below:
>
> Group A: (mm) 444 432 424 440 435 Average = **435** mm
>
> Group B: (mm) 441 460 410 424 440 Average = **435** mm
>
> **A**: Our results are better. They are all between 424 mm and 444 mm. Yours are spread between 410 mm and 460 mm.
>
> **B**: Our results are just as good as yours. Our average is the same as yours. We both got 435 mm for $d$.
>
> **C**: I think the results of group B are better than the results of group A.
>
> With which group do you most closely agree? Explain your choice.

Students' ideas were categorised using the scheme below:

***I agree with A because ...***

*UA01* - *Uncodeable*
*SA10* - *Smaller range/spread → mean more reliable/ more certain of result/mean*
*SA11* - *Smaller range/spread → mean more reliable; fewer outside factors*
*SA12* - *Smaller range/spread → mean more reliable; fewer errors/mistakes*
*SA13* - *Smaller range/spread → mean more reliable; group A more skilful/ careful*
*SA14* - *Smaller range/spread → mean closer to the exact/true distance*
*SA15* - *Smaller range/spread → mean more accurate*
*SA20* - *There is less deviance from the average / mean*
*SA21* - *Less deviance from mean → results better/more reliable because fewer external factors*
*SA22* - *Less dev from average → results better/more reliable because fewer errors/mistakes*
*SA23* - *Less dev from average → group A more careful/ skilful*
*SA24* - *Less deviance from average → results are closer to/ more consistent with true distance*
*SA25* - *Less deviance from average → results more accurate*
*PA30* - *One of the measurements (d = 435 mm) is equal/ identical to the average*
*PA35* - *One of the measurements (d = 435 mm) is equal to the average → more accurate*
*SA40* - *Results closer to together/ smaller range/ more reliable/ consistent or less uncertainty*
*SA41* - *Results closer together/ smaller range/ more reliable because less outside factors*
*SA42* - *A had less experimental error/ error bound/ mistakes than B (hence more reliable)*
*SA43* - *A performed experiment better/ more consistently than B (expt skill better)*

*SA44* - *Results of A are more consistent (closer together), closer to true distance/ reading*
*SA45* - *A's (individual) results are more accurate / more certain of accuracy of experiment*
*SA60* - *The result is more reliable*
*SA61* - *Result more reliable because fewer outside factors (better controlled)*
*SA62* - *A's result is less prone to/shows fewer errors/mistakes than B's*
*SA70* - *Results have a smaller standard deviation*
*SA73* - *A more careful in experiment → smaller standard deviation*
*SA74* - *Results have smaller standard deviation → greater chance of obtaining true mean*
*SA75* - *Results have smaller standard deviation → mean/ measurements more accurate*

***I agree with B because ...***
*PB01* - *Uncodeable*
*PB10* - *Measurements are more or less the same / makes sense to get similar answers*
*PB20* - *Most important to get the same average (no result better than another/ equally valid)*
*PB21* - *Same average most important- diff variables/conditions/outside factors caused variance*
*PB22* - *Same average most important - compensates for errors/mistakes in indiv. readings*
*PB23* - *Same average most important - range/spread/deviance from mean not important*
*PB25* - *Same average → equally accurate/ valid (variation expected)*
*PB30* - *Both did not obtain a recurring measurement (same value), average more important*
*PB40* - *Variation not important, more important to repeat many times*
*PB41* - *Different factors (for different experiments) could have affected results*
*PB65* - *Accuracy of individual results not under consideration - average important*
*UB70* - *Two intervals defined by ranges overlap, no reason that one is better than another*

***I agree with C because ...***
*PC01* - *Uncodeable*
*PC40* - *B's results vary more*
*PC41* - *B shows greater variation in results which more clearly demonstrates that external factors affect experiments*

*Choice A: Point paradigm*

The readings of group A (but not group B) included one datum that was identical to the average. Code categories PA30-35 indicate reasoning that an experimental result is superior when the mean is identical to an individual reading or readings. Respondents here do not acknowledge the spread in data at all and believe that single measurements are the sole indicators of experimental expertise and success. A student gave the following response:

> *The correct result must be 435, therefore Group A's results are more accurate.* (PA35)

*Choice A: Set paradigm*

A narrower spread in data may be seen to indicate that the calculated mean is more reliable, thus implying that the results are better (SA10-15). Examples of mainstream students' explanations are given below:

*Group A's readings are more consistent and will end up with less uncertainty in their final answer.* (SA10)

*The values calculated for d in group A do vary over a smaller range which shows that their mean calculated is more precise and more accurate compared to group B.* (SA15p)

Students may have argued that the individual readings of group A show less deviance from the mean than for Group B (SA20). They may further have reasoned that this implies that group A had fewer external factors to contend with (SA21), made fewer errors or mistakes (SA22), were more skilful in conducting the experiment (SA23), obtained a result closer to the 'true value' (SA24) or that group A's results are more accurate (SA25). The response quotes below are representative for this cohort:

*Group A's results show a stronger tendency toward the mean and are not as spread out. This suggests that they are better.* (SA20)

*In an ideal world a value close to 435 would have been obtained every time. Group A has less deviation from the ideal, so has minimised error more effectively. Their result (even though the same) can be trusted more.* (SA22)

*Group A had much more consistent readings than Group B, therefore their results are much closer to the average distance of group B and therefore more accurate than Group B.* (SA25)

Group A's results may have been perceived to be superior based on the observation that individual readings are closer to each other numerically (SA40-45). For these arguments, the calculated means are not mentioned. The responses below were typical for these code categories:

*What if the readings are spread between 1 to 870? It would be terrible. The closer the numbers are, the better. Otherwise, one might think there is something wrong with the experiment.* (SA40)

*Simply because the results are all so close together (difference between highest and lowest is 2cm). This indicates that other variables were minimised (more so than group B), thus the results weren't as greatly affected (and therefore, more accurate). However, it is possible that both experiments were carried out as well as each other, just the reliability of the equipment may have differed.* (SA41a)

*The spread of readings obtained by Group A cover a smaller interval than those of Group B. The readings of group A were more consistent, closer to true reading.* (SA44)

*The margin of error for A is smaller than for B, which suggest that A is more reliable. The numerical equivalency of the averages is more a matter of coincidence than accurate measurement, as the wildly varying readings could easily have provided a far different reading, especially if release 2 or release 3 were excluded form the calculation.* (SA42a)

*If I didn't look at the "final average measurement", I would trust the accuracy of group A more that that of group B, since group A's measurements are clustered very closely together - suggesting that they work very accurately (more so than group B!). That the "final results" (d=435) are the same could be due to chance.* (SA43a)

*The less range there is with a set of answers, the more accurate that set is. Therefore Group A's answers are more accurate because there is less deviation therefore Group A's answers are better.* (SA45)

Code categories SA60-62 were assigned when responses used the reliability of "the result" as the basis for the argument. It is not clear from the responses what respondents meant "the result" to be. The response below is representative:

*Group A's result shows little errors as compared to Group B's result.* (SA62)

Students may have recognised that the results of one group have a smaller standard deviation, the statistical measure of spread present in a data set of multiple repeated measurements (SA70). They may further have argued that the group with the smaller standard deviation was more skilful in conducting the experiment (SA73), or would have had a greater chance of obtaining the true mean (SA74) or that the smaller standard deviation implied that the readings and/ or mean were more accurate (SA75). The small number of responses for this cohort of students are shown below:

*The smaller the range over which the data is spread, the better the data. This would be clear if the standard deviation of the average for each group was calculated. The standard deviation of A would be much smaller than that of B.* (SA70)

*The calculated deviation for group A will be less than group B. Which means for the same confidence level group A could make more precise predictions.* (SA70p)

*A smaller spread of the distance readings of group A than group B indicates a higher precision in carrying out the experiment. That is, more care was probably taken. This will result in a smaller standard deviation of the mean.* (SA73pm)

*Assuming d=435 to be the correct average, Group A had more chance of finding it, since their results show a smaller standard deviation.* (SA74)

*A's value for d is indeed better, because the measurements all lie within a much smaller bound than for B. Thus, the standard deviation for A (a measure of the uncertainty in d) will be lower than B, and thus A is likely to be more accurate* $\{\sigma_A \approx 7\ \sigma_B \approx 18\}$. (SA75)

There were no responses from this cohort that could not be categorised within the point and set framework (UA01).

*Choice B: Point paradigm*

This procedural decision is indicative of the notion that the mean is the sole outcome of an experiment involving repeated measurements. Differing spread in data is viewed as incidental and of no importance. Respondents may have judged the individual results of the two groups to be similar and hence expected the calculated means to be the same (PB10), as one student reasoned:

*Both group A and B used proper experimental procedure and, as a result, obtained similar results. Thus both results appear equally genuine.* (PB10)

Another view is that the mean is the only important information that can be extracted from results. The spread is therefore not recognised as an integral part of the measurement result (PB20). The quotes below include those where students reasoned that the variation in results was due to differing outside factors (PB21), due to "errors" or mistakes and therefore the mean compensates for these (PB22; none so coded for this cohort), or that the spread is not important (PB23) or that the identical means imply that both groups were equally accurate (PB25):

*They may vary in ranges, but the mean is the same, therefore both groups' results are equally valid.* (PB20)

*The final result, the average is the important information from the data. Different situation cause different results but the average should remain constant.* (PB21)

*As we speak of the accuracy of results, they are equally good as they come to the same result. If one speak of the manner in which the readings were made, then Group A has readings that were closer to the mean, thus need less tries to come to an unbiased reading.* (PB23a)

*Both final answers reflected the situation accurately. It is impossible to get identical results for an experiment without doing the experiment hundreds or even thousands of times.* (PB25)

Individuals may have persisted with reasoning used for the data collection and UR probes that argued for a recurring value to represent a measurement result. As the results for both groups do not contain a recurring number, the identical means are seen to be the important consideration (PB30). A student wrote the following explanation for siding with B:

> *They both did not get the same value for the five releases they made and the average is the most important out of all the other values.* (PB30)

Students may simply have viewed the variation in readings as not important and believed that repetition is more important (PB40) or that the variation was caused by different factors (PB41). No direct reference is made to the means or individual readings as in the response quotes below:

> *Because it is not all that important the difference between the highest distance and lowest. It is more important to perform the experiment several times.* (PB40)

> *The varying of results does pose some problems, but only when working with the error. Other than that the results were from different experiments and would have had different factors working for or against them.* (PB41)

A student might have stated that the "accuracy" of individual readings is not sought for and hence viewed the calculated mean as the only way to represent results, as below:

> *The accuracy of each result is not what is being looked at. The average is the important thing and therefore both are equally good.* (PB65)

*Choice B: Undecided paradigm*

A rare response type is when a student takes a procedural action associated with the point paradigm but justifies this with a set reason. The fact that the intervals of the two data sets overlap may be argued to imply that neither measurement is superior (UB70). Data set quality is convoluted or confused with comparison. The quote below illustrates this reasoning:

> *Besides the fact that the two averages are the same the range of values of group A is from 424→440 and the range of values of group B is 410→460 so we can see that these two intervals overlap so there is no reason to say than one measurement is better than the other.* (UB70)

*Choice C: Point paradigm*

Students either stated that group B's results vary more and is therefore better (PC40) or that the greater variation in the readings made by group B better demonstrates the effect of outside factors on experimental results (PC41). Spread is not used as an indicator of data set quality. No responses were coded PC01. The following response was given:

> *Group B's results demonstrate that external conditions do in fact play a major role in this experiment. Their results therefore give a clearer indication of these effects therefore their results are better.* (PC41)

In Table 3.6, the response categories that were assigned for this cohort pre- and post-course are grouped and described. The table summarises how the mainstream students dealt with the aspect of data set quality.

*Table 3.6: Summary and frequency of code categories established for mainstream students' responses to the SMDS probe before and after instruction. (n =53)*

| Codes | Descriptor | Pre-course | Post-course |
|---|---|---|---|
| PA30-35 | Results are better if individual reading/s are equal to the mean. | 1 | 0 |
| PB10 | Readings judged to be more or less the same; hence means are the same, which implies the same quality of measurement. | 2 | 0 |
| PB20-25 | Most important consideration for means to be the same (despite different ranges) to judge quality. | 11 | 8 |
| PB30 | In the absence of a recurring reading, the mean is the only way to judge the quality (equal) of a data set. | 1 | 0 |
| PB40-41 | The variation in individual reading are not important when considering the relative quality of data sets (the mean only). | 1 | 1 |
| PB65 | Accuracy of individual readings not important, mean most important for making quality judgement. | 1 | 0 |
| PC40-41 | A data set with a greater spread is better. | 1 | 0 |
| SA10-15 | The data set with the smaller spread has a more reliable mean. | 3 | 2 |
| SA20-25 | A smaller spread shows that the better result's readings deviate less from the mean value. | 7 | 4 |
| SA40-45 | Data sets with readings closer together are more reliable/ better. | 22 | 17 |
| SA60-62 | The "result" is more reliable/ better (clearly based on smaller spread) | 1 | 0 |
| SA70-75 | A data set with a smaller spread has a smaller standard deviation, hence is the better result. | 2 | 20 |
| UB70 | Data sets with overlapping intervals (+equal means) are equally good; no one better than the other. | 0 | 1 |
| | **Total** | **53** | **53** |

### 3.3.7 Comparing data sets with differing means and similar spreads (DMSS)

The DMSS probe asks respondents to decide whether or not the results of measurements made by two groups of students agree with each other. Value-by-value comparison, whether individual readings or range sizes, of the two data sets and judgements based on the perceived proximity of the means are associated with the point paradigm. Where the degree of overlap of the intervals of the two groups only is used as a basis for the compatibility of two measurements, the set paradigm is indicated. The probe contains the following text:

Two other groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm. Their results for five releases are shown below.

| | | | | | | |
|---|---|---|---|---|---|---|
| Group A: (mm) | 440 | 438 | 433 | 422 | 432 | Average = **433** mm |
| Group B: (mm) | 432 | 444 | 426 | 433 | 440 | Average = **435** mm |

**A**: Our result agrees with yours.

**B**: No, your result does not agree with ours.

With which group do you most closely agree? Explain your choice.

Students' ideas were categorised using the coding scheme below:

***I agree with A because ...***

*UA00* - *No reason*
*UA01* - *Uncodeable*
*PA10* - *Means similar + readings lie in narrow range/error bound/closely centred around means (similar)*
*PA20* - *Small difference between averages*
*PA21* - *Small difference between averages: difference due to external factors*
*PA22* - *Small difference between averages: difference due to/within experimental errors*
*PA23* - *Small difference between averages: diff due to systematic uncertainty of equipment*
*PA24* - *Small difference between averages → both close to true value*
*PA25* - *Small difference between averages → both groups equally accurate*
*PA26* - *Small diff between averages can be ignored as due to minor statistical differences*
*PA27* - *Mean and std. deviation of both d's roughly the same*
*PA28* - *Small difference in averages; this diff would minimise with more measurements*
*PA30* - *Two groups have readings that are identical &/or similar, diff in average negligible*
*PA33* - *Both did not obtain identical values for all 5 releases and diff in averages small*
*PA40* *Small difference between averages with similar/nearly same ranges/spreads*
*PA42* - *Expt error reason for small difference in range, negligible as averages almost equal*
*PA43* - *Difference in average is small/ negligible relative to large deviation in readings*
*PA45* - *Small difference between averages with similar spreads, therefore both accurate*
*PA60* - *Averages are only estimates → averages agree since approximately equal*
*PA63* - *Probable that averages are similar and not identical since the individual results deviate from average randomly*
*PA65* - *Results are both averages → no large degree of accuracy required/expected*
*SA70* - *Intervals/ ranges/ uncertainties overlap(readings from one group fall in other's range)*
*PA71* - *Average falls within range/spread/uncertainty of other group's result/s*
*SA72* - *Two sets of results agree to within experimental error as their intervals overlap*
*PA80* - *If the most wayward measurement is excluded, results/averages agree better*

***I agree with B because...***

*PB00* - *No reason*
*PB01* - *Uncodeable*
*PB10* - *A/B is better - have more individual readings closer to the average*
*PB15* - *Degree of accuracy of B is better, gap between consecutive readings smaller cf A*
*PB20* - *Averages are different (even though close)*
*PB21* - *Averages are different due to different conditions/external factors*
*PB22* - *Averages are different due to experimental errors*
*PB23* - *Averages differ even though average lies in range of other group's measurements*
*PB25* - *Averages are different - absolute accuracy is required to agree*
*PB26* - *Averages different – no expt uncertainty given / std deviations might be different*
*PB28* - *Averages are different – if more readings were taken they could agree*
*PB30* - *There are no clear/ common similarities/ recurring readings between two sets*
*PB33* - *Small difference in averages but large difference in individual measurements*
*PB40* - *Averages are different and ranges/spreads are different*
*PB60* - *They simply don't agree*
*PB61* - *Each group completed experiment under different conditions/ ext factors*
*UB73* - *Least count on average used as uncertainty interval, thus mutually exclusive*
*PB80* - *A/B better-if most wayward measurement is excluded, results/averages agree better*
*PB90* - *Last digit is uncertain, so rounding off will result in A=430 and B=440, so different*

*Choice A: Point paradigm*

Respondents who side with group A judge that the means are nearly identical or "close together". Students may have argued further that the individual readings for both groups are scattered closely around the respective means (PA10). They may have believed that the similar quality of two data sets is an important consideration in deciding whether two results agree or not, as in the response quote below:

> *Although the means of the 2 groups are not equal, they are very close in value. If one looks at both sets of data; both are quite closely centred around their means.* (PA10)

Arguments could be based primarily on the "closeness" of the means (PA20). In addition, various explanations for this difference in the means may have been offered; due to external factors (PA21), experimental "error" (PA22), "systematic uncertainty" of equipment (PA23) or minor statistical differences (PA26). Respondents may have been convinced that the "closeness" of the means implies that both results (i.e. means) are close to the "true" value (PA24) or that both are equally accurate (PA25). Others may have thought that the means and standard deviations are similar (PA27) or that more measurements would have minimised the difference in the means (PA28). The sample quotes below typify those for the code categories described above:

> *The results may not be exactly the same - but the difference is 2mm which, when measuring the distance of a ball is not a great difference. Hence, the two answers agree with each other in the respect that they*

*are almost the same and not hugely different. In some other experiments however, 2mm can make a big difference in the results and they would not agree with each other.* (PA20)

*For an experiment like this one where many other factors (e.g. smoothness of slope, roundness of ball) can easily cause errors greater than 2mm, I think the 2 groups' results do agree.* (PA21)

*I agree with A that the averages agree because their average is -2mm of B. Experimental error, different situations and different people handling the equipment lead to slightly different readings. They do not have to be equal to agree.* (PA22x)

*There is a slight difference between the two averages. This is because every instrument has its own uncertainty or a slight error. Therefore the results are the same.* (PA23)

*The results are likely to be different from each other because it's very difficult or almost impossible to obtain the "perfect" answer as some factors may vary a little with each experiment. The results agree, because they are close to each other, and probably close to the "perfect" answer.* (PA24x)

*The difference between the two averages is only 2 mm, which is an acceptable margin of error. Also, if the two averages are averaged, the result would lie halfway between the two, suggesting that both are fairly accurate.* (PA25)

*Although the results are different this difference is so slight (± 0.5 %) that we can ignore it and see the results as minor statistical differences; the experimental results agree.* (PA26)

*The standard deviation and the mean of both d's are roughly the same.* (PA27)

*The averages are very similar so there is no large conflict in results. There is a good chance that the difference in averages would minimise with more measurements.* (PA28)

Arguments may be centred on whether some or all the individual readings are identical for both groups. Responses either indicated that the two groups have common readings with calculated means which are not very different (PA30) or observed that neither group produced recurring readings and show only a small difference in their means (PA33), as in the quotes below:

*The two groups both obtained values d=440, d=433, and d=432. Where the results differ, this difference is minimal. The difference between the averages is less than 0,5% of the total.* (PA30)

*The averages obtained seem to compliment each other. They are only out by 2mm, and one should bear in mind that different apparatus was used, which could be responsible for this small change. In addition, some readings were in common with both groups.* (PA30x)

*They both did not get the same value for d for the five releases they made and there is a very small difference between their average values.* (PA33)

Where the decision was based on whether the means are in close proximity of each other in combination with how similar the two ranges of readings was believed to be, code categories PA40-45 were assigned. Results may have been deemed compatible where both the means and spreads were thought to be similar (PA40; PA45 for concluding that both groups are accurate). Others may have believed that the "small" difference in means compensates for or even negates the deviation observed in the data (PA43; PA42 if deviation was believed to be caused by "experimental error"). The quotes below are representative:

*The average values calculated are close enough to one another and the range over which the results of d spans for each group is more or less similar; both are not huge ranges and lie close to one another i.e. group A from 422-440 and grp B from 426-444.* (PA40)

*A's results range from 422 to 440 and B's results range from 426 to 440, which means that A has made a little error (negligible since averages are almost equal) as compared to B.* (PA42)

*When there is a deviation of 18mm in both the two sets of answers, a difference of 2 in the average is negligible.* (PA43)

*The averages are more or less the same, therefore it can be said the results are in agreement and the range of A is 422-440 and the range of B is 426-444 which shows accuracy and reliable results from both groups.* (PA45)

Students may have justified why they believed that the perceived small difference in the means is sufficient to make a decision on the compatibility of data sets. Some could have viewed the average as an estimate and hence concluded that the data sets agree based on the means being approximately equal (PA60). Others may have stated that they expect a small variation in the calculated means based on the fact that the individual readings deviate randomly from the mean (PA63). Yet others could have argued that a high degree of accuracy is not required or expected in this experiment (PA65). The responses below were received from this cohort:

*It is an average of the results, which is only an estimate and 433 mm ≈ 435 mm.* (PA60)

*As the individual results of each release deviate from the average randomly, it is probable that averages from different groups would be similar, but not identical.* (PA63)

*The results arrived at are both averages, therefore no great amount of accuracy can be expected. None of Group B's measurements were exactly 435 therefore they can't claim that the correct answer is 435. All they can say is that it is between 432 and 444 and close to 435.* (PA65)

Students may have judged that the mean of one group falls within the uncertainty interval of the other group's results (PA71) as was the case in the following response:

*Both means lie within the standard deviation of the other group's data.* (PA71σ)

Some respondents may have argued that the results, i.e. means, would have agreed better if the most wayward readings were excluded (PA80). The response below illustrates this category:

*As we saw earlier, it was possible for the ball to reach 588mm therefore if one were to exclude again the longest distance the two would be virtually the same, i.e. exclude the 444mm and then the results to a degree, agree with each other.* (PA80)

<u>*Choice A: Set paradigm*</u>

The following responses represent those where the overlap of the two group's ranges or intervals were used to argue that the two groups' results agree (SA70). Some may have mentioned experimental error (SA72) while others may have explicitly defined the intervals as the mean together with the statistical uncertainty (standard deviation/ of mean):

*Both readings have a corresponding uncertainty: thus the answers each group obtains should be considered more as a range of values than one exact value. These ranges intersect, so the results do agree.* (SA70)

*From the standard deviation of the means we can see that the measurements of group A lies in 426-439 and the measurements of B lies on 429-441 so we can see that these intervals overlap, therefore the two measurements agree.* (SA70m)

*Group A obtained answers between 422 and 440, B between 426 and 444. These ranges overlap. When the values for d are expressed as an interval, i.e. mean± standard deviation, the two groups' answers will probably still coincide.* (SA70σ)

*Choice A: Undetermined paradigm*

Students may have given uncodeable reasons for siding with A and hence could not be classified as either point or set. There were no such responses for this cohort of students.

*Choice B: Point paradigm*

Not a single response defending choice B, from diverse groups of students, could be associated with the set paradigm. Hence, it was appropriate to view responses that did not provide a written argument or where the reason could not be categorised, as consistent with the point paradigm. No responses from the mainstream students of this study were coded PB00 (no reason) or PB01 (not codeable).

Students could have based their decision on the erroneous observation that the readings of one group is more closely scattered around the respective mean (PB10). This again is indicative of the confusion between data quality and compatibility. Others may have considered the difference in consecutive results and concluded that one group is more accurate (PB15). The following responses were received:

> *Group A's readings are better because they have more readings closer to the average than group B and therefore their readings don't really agree with each other.* (PB10)

> *The degree of experimental accuracy of B is higher since the biggest gap between experimental results is only 6 while in A it is 11.* (PB15)

The ideas identified for code categories PB20-28 are very similar to those for PA20-28, except here the proximity of the means is deemed to be not sufficient to conclude that the data sets agree. The sample responses below are representative:

> *The two results don't agree in that they have different answers. To agree, they should have exactly the same readings/ answer.* (PB20)

> *Can only agree if difference between the averages is smaller than one.* (PB20)

> *The final values do not necessarily agree because there may have been errors that arised when the experiment was done.* (PB22)

*The average for both groups differ hence they do not agree with the other, even though their average is in the same vicinity of the other group's measurements.* (PB23)

*Scientifically speaking, there is a difference between the results, being 2 mm. So if accuracy is what counts, the results do not agree.* (PB25)

*Strictly speaking, group B's average is higher than group A's, therefore the distribution of B's reading is slightly shifted to a longer distance. Besides, the standard deviations might be different.* (PB26σ)

*The average value for d is different for each group, even though they differ by only 2 mm, and therefore their readings do not agree. Perhaps if more readings were taken the results might become the same, but until that can be shown the readings are different.* (PB28)

Students may have made a point-by-point comparison of the individual results obtained by the two groups and concluded that the results cannot agree since there is no clear common or recurring reading between the two data sets (PB30). Alternatively, it may have been recognised that even though the means are relatively similar, individual readings are quite different. The following responses were given:

*There is no clear connection between all of the values from each group. There are similarities, but there is not a common similarity.* (PB30)

*For the results to agree, the results should be as close as possible if not the same. Although the average had only a difference of 2 mm, there was a big difference in the two largest measurements. Thus I feel the results are not agreeable enough.* (PB33)

The respondents might have thought that for compatibility two data sets must have both identical means and data ranges (PB40). The following quote is an example:

*The average is different and the range of answers is different. Group A's answer range between 440 and 422 while group B's range between 444 and 426.* (PB40)

Code categories PB60 and PB61 were assigned in instances where the written text did not mention the means, spread or individual results explicitly, yet the reasoning clearly implied that such one-to-one comparisons were made as below:

*Because they simply don't agree with each other. (PB60)*

*There may have been slight differences between the conditions under which each group completed the experiment. (PB61)*

Students may have argued that the means would have been nearly identical if the most extreme measurement was excluded (PB80). Another approach espoused the use of some rule to round off the means; the uncertainty associated with the last digit of the mean and hence rounded up or down which yields means 10 millimetres apart (PB90). Sample quotes for these two code categories are presented below:

*I mistrust the measurements made by group A, although the results of both groups range over 18 millimeters. Whereas group B's measurements are 7 mm out (from the nearest other measurement) at most, there is a 10 mm difference between group A's smallest result (422 mm) and their next one up (432 mm). They could have disregarded this result (422 mm) and ended up with an average of 435.75 – nearly exactly the same as group B's result.* (PB80)

*Because there is no decimal place, we have to assume that the last digit is uncertain. But if you rounded 433, you'll get 430, whereas if you round 435, you get 440. The two are different.* (PB90)

*Choice B: Undetermined paradigm*

A concept, which is often introduced at first year level, is the least count of measurement apparatus. Students could have erroneously applied this method to the two means and hence defined the intervals of uncertainty as the mean plus or minus half a millimetre. Then the appropriate conclusion is that the two intervals do not overlap, which is reasoning associated with the set paradigm. However, disregarding the spread in a data set as an integral part of the result is an action associated with the point paradigm. UB73 was assigned for the responses below:

*The uncertainty in the readings is 0.5 mm so the average of A's data set lies in the range (432.5;433.5) and B's average lies in (434.5;435.5). These intervals are mutually exclusive and indicate a disagreement in the findings of group A and group B.* (UB73)

*Since the metre rule is calibrated in units of 0.1 cm, the standard deviation of the average values should be 0.1 cm. The difference in averages here is 0.2 cm, hence the answers do not agree. (UB73σ)*

The range of response categories, demonstrating how mainstream students dealt with the aspect of data compatibility, are grouped and described in Table 3.7.

*Table 3.7: Summary and frequency of code categories established for mainstream students' responses to the DMSS probe before and after instruction. (n =53)*

| Codes | Descriptor | Pre-course | Post-course |
|---|---|---|---|
| PA10 | Data sets with "similar" means *and* readings closely centred around means are compatible. | 0 | 3 |
| PA20-28 | Compatibility based on "closeness" of means only/ small differences expected due to various experimental factors. | 29 | 15 |
| PA30-33 | Agreement if small difference in means & a few identical readings. // Alternatively, in absence of recurring datum, "similar" means are enough to have agreement. | 4 | 5 |
| PA40-45 | Agreement based on results having "similar" means and spreads agree // "small" difference in means negates large spread in readings. | 6 | 4 |
| PA60-64 | Agreement based on means seen as estimates, "small" difference caused by randomness, or high degree of accuracy (equal means) not required. | 2 | 0 |
| PA71 | Results are in agreement if the mean of one group falls within uncertainty interval defined by the other groups' readings. | 0 | 3 |
| PA80 | Means would be closer together or equal if most wayward readings are discarded. | 1 | 0 |
| PB10 | The data set, which contains readings that are closer to the mean or differ less for consecutive repeats, is better. | 1 | 0 |
| PB20-28 | No agreement unless means are identical. | 6 | 9 |
| PB30-33 | Only data sets with clear recurring/ identical readings between them are compatible. | 1 | 1 |
| PB40 | Both the means and data ranges need to be identical for compatibility. | 1 | 2 |
| PB60-61 | "They don't agree" (differing means & dispersion must be due to different experimental conditions experienced when collecting data) | 1 | 0 |
| PB80 | Agreement requires identical results, however discarding the most wayward reading/s of either data set might yield more equal means. | 1 | 0 |
| PB90 | Data sets are compatible if the rounded off mean values are identical. | 0 | 1 |
| SA70-72 | Data sets are compatible when the uncertainty intervals (defined by the data ranges) overlap in some region. // Some readings from one set fall within the other's range. | 0 | 8 |
| UB73 | Agreement if the uncertainty intervals defined by the least count of the measuring apparatus, overlaps. | 0 | 2 |
| | **Total** | **53** | **53** |

### 3.3.8 Comparing data sets with differing means and overlapping spreads (DMOS)

The DMOS probe is similar to the DMSS probe. Respondents need to consider whether two data sets are in agreement based on whether their ranges or uncertainty intervals overlap. The actions and reasoning behind decisions are thus similar to those described for the DMSS probe. The text which appears on the DMOS probe sheet follows:

Two groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm. Their results for five releases are shown below.

| | | | | | | |
|---|---|---|---|---|---|---|
| Group A: (mm) | 444 | 435 | 424 | 440 | 432 | Average = **435** mm |
| Group B: (mm) | 458 | 438 | 462 | 449 | 443 | Average = **450** mm |

**A**: Our results agree with yours.

**B**: No, your results do not agree with ours.

With which group do you most closely agree? Explain your choice. Do not use the word "results" in your explanation.

The coding scheme used to categorise responses to the DMOS probe is presented below:

***I agree with A because ...***

*UA00* - *No reason*
*UA01* - *Uncodeable*
*SA10* - *Smallest readings of B and largest readings of A are similar / some readings of A/B fall within the range of B/A*
*PA20* - *Difference in means (averages) small (the % difference is acceptable)*
*PA40* - *Same spread/ range (comparison between the widths of two spreads)*
*SA70* - *Intervals defined by uncertainties (probably) overlap*

***I agree with B because...***

*UB00* - *No reason*
*UB01* - *Uncodeable*
*PB10* - *Individual measurements differ greatly / general spread/distribution of B higher*
*PB11* - *Only two values of similar size, rest very different*
*PB20* - *Difference in means / averages too large*
*PB21* - *Diff in means large; individual releases/measurements for B higher than A/not close*
*PB22* - *Diff in means large; ranges/ general distribution of readings differ*
*PB23* - *Means / averages are different, even though ranges are of similar size (same spread)*
*PB24* - *Means / averages largely different → uncertain of true value for d*
*PB26* - *Means / averages different: can't be compared further since no uncertainties given*
*PB27* - *Means / averages different: not in same vicinity/ range of other group's measurements*
*PB30* - *No common/identical individual results between two data sets*
*PB40* - *Ranges / spread different*
*PB41* - *Ranges are different, although spread of data similar (+ only 1/2/3 readings common to both ranges)*
*PB43* - *A has a smaller uncertainty/ from mean than B*
*PB61* - *Apparatus / settings must have been different for both groups, hence difference*
*SB70* - *Intervals/ranges defined by uncertainties do not overlap*
*PB71* - *Means do not lie within uncertainty interval of other group*
*UB73* - *Uncertainty interval defined by least count (1 mm/0.5 mm) → do not overlap*
*SB75* - *Intervals defined by mean +/- uncertainty do not overlap; both equally accurate since both uncertainties are small*

*Choice A: Point paradigm*

Respondents may have judged the difference in the calculated means to be small enough for the results to agree (PA20) or gauged that the two data ranges are similar in size (PA40). The responses below were given:

*The ratio* $\frac{450 - 435}{450}$ *is a small number. Because d is a big number, one should not be surprised that both readings deviate slightly.* (PA20)

*They have the same spread, comparing their smallest & largest values, they are both the same.* (PA40)

*Choice A: Set paradigm*

Respondents may have based their decision on the observation that there are individual results which fall within the ranges of both data sets (SA10). It is recognised that a measurement result is defined by an interval, even though the formal construct of uncertainty is not used. There were no responses of this type given by this cohort. In Section 3.1.3 it was noted that the intervals of the two results defined by the mean and one standard uncertainty (standard deviation of the mean) do not overlap [$d_A$ = 435.0 ± 3.4 mm (68% confidence interval); $d_B$ = 450 ± 4.5 mm (68% CI)]. However, students' action and reasoning were deemed to be consistent with the set paradigm when they considered intervals defined by one standard deviation ($\sigma_A$ = 7.7 mm; $\sigma_B$ = 10 mm), and hence came to the conclusion that the results agree (SA70). One student gave the following written justification:

*The obtained values for d do agree. The uncertainty for both averages is relatively small – between 7 and 10, but the intervals in which the respective values may lie do overlap, and hence the results agree.* (SA70)

*Choice A: Undetermined paradigm*

Not giving a reason (UA00) or providing written text that cannot be categorised within the scheme (UA01) does not place responses unambiguously within the point or set paradigms. This cohort of students did not respond to the DMOS probe in these ways.

*Choice B: Point paradigm*

Value-by-value comparison of the two groups' readings may have lead students to conclude that the values differ and hence the two data sets are incompatible (PB10). Alternatively, respondents may have observed that only one or two individual results across the two data sets are similar, with the rest of the readings differing substantially between sets (PB11). The two quotes below illustrates these response types:

*Because for each release, group B always obtain a value higher than group A; group B's general spread is higher than group A's. So they don't agree with each other.* (PB10)

*If you look at the individual results rather than the mean value, there are only 2 values in each group of similar size the rest are very different.* (PB11)

The series of response types coded PB20-27 focuses on the judgement that the means calculated are too far apart (PB20). Some students may have included secondary reasons to bolster their arguments; the individual readings are different (PB21), the data ranges differ (PB22), the variance in the means causes uncertainty in what the true value of *d* should be (PB24) or the means do not fall in the other group's range of readings (PB27). The primary argument may have centred on the perceived large difference in means but qualified with the observation that the ranges or spread in data are of a similar size (PB23). Some students may have argued that in the absence of calculated uncertainties, the difference in the means is the only measure for comparing the two data sets (PB26). The responses below represent the categories discussed above:

*There is a significant difference between the two groups' conclusions. In an experiment of this sort and on this scale, it should be said that the separate experiments do not agree.* (PB20)

*450 mm is too far from 435, I think that there is a common error in the experiment. Maybe the ramp is more inclined or there might be an external force every time they release the ball.* (PB20)

*The readings that group B took were generally much higher and so was their average, the results are 'drastically' different and do not agree.* (PB21)

*There is a much larger difference between the data now, and it is too great. B also has a much larger range than A hence the large variability of the average d.* (PB22)

*The averages lie too far apart to 'agree' with one another, even though the ranges for each set of values is of a similar size. Each group has an accurate mean value but one group doesn't have a correct answer or had a factor which constantly affected each measurement.* (PB23xa)

*The average value for d is not the same and can't be compared any better because no uncertainty is given.* (PB26)

*The averages are not the same and are even not in the same vicinity as the other measurements.* (PB27)

Students may have insisted that two data sets contain common individual results for them to agree (PB30); this cohort of students gave no such responses. Answers could have been based on the difference in the data ranges or spreads (PB40) or uncertainty intervals (PB43). Some may have argued that in spite of the sizes of the spreads and or some readings being similar, the differing ranges imply incompatibility (PB41). The variance in the means and spreads may have lead students to conclude that the two groups must have performed the experiment under different experimental conditions (PB61). The following responses represent the categories described above:

> *The conclusion reached by each group do not agree. Group A has a lower standard deviation from mean than Group B, and group B has a larger range of values.* (PB43σ)

> *It is obvious that the apparatus or setting was different for the two groups.* (PB61)

Certain students may have applied the test of overlapping intervals to the means only, and argued that if the mean of one data set does not lie in the other groups' uncertainty interval, compatibility is disproved (PB71). The following quotes typify this kind of reasoning:

> *The range of uncertainty in A does not extend to the average value calculated in B. The same applies for B – the lowest value for B is 438 mm whereas the value obtained for A is even lower than that.* (PB71)

> *It is hard to say, but from sight alone I would suggest that the standard deviation of group B's readings is just less than 15. Thus, 435 does not lie within their value range and the results do not agree (450 certainly does not lie in the range for group A's mean).* (PB71σ)

<u>*Choice B: Set paradigm*</u>

It may have been determined that the two uncertainty intervals do not overlap (SB70) and further argued that the results are equally accurate since the uncertainties are similarly small in size (SB75). Some sample responses follow below:

> *The answers obtained by the two groups have a corresponding uncertainty; thus they should rather be considered as a range of values. These ranges do no intersect, thus the answers obtained by the groups do not agree.* (SB70)

> *The regions bounded by the mean ± standard deviation which the two groups calculated do not intersect.* (SB70σ)

*The two averages together with their std. deviations do not overlap anywhere. The small std. deviation of both averages tells me that both experiments are quite accurately done.* (SB75σ)

*Choice B: Undetermined paradigm*

Siding with B do not necessarily place responses in the point or set paradigm, hence the need for the UB00 (no justification given) and UB01 (written text cannot be categorised) assignments. As explained before (category UB73, DMSS probe), using the idea of a least count to define uncertainty intervals and hence conclude that the results do not agree (UB73) represents an action consistent with the point paradigm and reasoning consistent with the set paradigm. One student gave the following response:

*Since the uncertainty in the readings is 0.5mm the ranges of averages for A and B are (434.5; 455.5) and (449.5; 450.5). Since these intervals are mutually exclusive, the findings disagree.* (UB73)

Table 3.8 shows the range of code categories with descriptions, identified for the mainstream students after completing their laboratory course on data analysis. The table presents the broad ideas these students used when considering data set compatibility.

*Table 3.8: Summary and frequency of code categories established for mainstream students' responses to the DMOS probe after instruction. (n =53)*

| Codes | Descriptor | Post-course |
|---|---|---|
| PA20 | The difference in the means is small enough, hence data sets agree. | 1 |
| PA40 | The ranges of the two sets of data are similar in size, hence the results agree. | 1 |
| PB10-11 | The individual readings between sets differ greatly, hence disagree. | 3 |
| PB20-27 | Difference in means too great for agreement. | 28 |
| PB40-43 | The data ranges are different, hence no agreement. | 4 |
| PB61 | The variance in means and spreads is due to experiment being performed under differing conditions, hence cannot agree. | 1 |
| PB71 | The mean of one data set must fall within the uncertainty interval of the other for compatibility. | 6 |
| SA70 | Data sets are compatible, since the uncertainty intervals overlap in some range or, if some readings fall in the other's range of values. | 2 |
| SB70/75 | Data sets are not compatible, as the uncertainty intervals (one std dev of mean) do not overlap in some range. | 6 |
| UB73 | No agreement since uncertainty intervals defined by least count, do not overlap. | 1 |
| | **Total** | **53** |

### 3.3.9 Comparing data sets with differing means with uncertainties equal in size (DMSU)

The DMSU probe asks respondents to compare two results presented in the formal manner, calculated means together with their uncertainties (standard deviation of the mean), instead of a series of individual measurements for several releases. The actions and underlying reasons used by students to decide on the agreement of the two data sets are thus similar to the DMSS and DMOS probes. The text used on the DMSU probe sheet is below:

> Two other groups of students compare their results for $d$ obtained by releasing the ball at $h$ = 400 mm. Their means and standard deviations of the means for their releases are shown below.
>
> Group A: $d$ = 436 ± 5 mm
>
> Group B: $d$ = 442 ± 5 mm
>
> **A**: Our results agree with yours.
>
> **B**: No, your result does not agree with ours.
>
> With which group do you most closely agree? Explain your choice. Do not use the word "result" in your explanation.

The coding scheme follows:

***I agree with A because ...***

*UA00* - *No reason*
*UA01* - *Uncodeable*
*PA10* - *The standard deviations are exactly the same*
*PA13* - *The std deviations are the same; implies variation in values are the same*
*PA14* - *The std deviations are large; implies both far from true value for d (both wrong)*
*PA20* - *Means are similar (difference small) + standard deviations exactly the same*
*PA21* - *Std dev's same + small difference in means caused by different apparatus/ conditions*
*PA23* - *A's/B's maximum/minimum mean value close to B's/A's calculated mean*
*PA27* - *Means very close; not important for means to lie in other groups' uncertainty intervals*
*SA40* - *Values for d agree to within experimental uncertainty*
*PA60* - *The results are nearly the same*
*SA70* - *Intervals, defined by means and standard deviations, overlap (in certain range)*
*SA72* - *Intervals overlap in a range (437-441): A / B could have under- / over-measured*
*SA74* - *Intervals overlap, probable that true value lies in combined error/confidence interval*

***I agree with B because...***

*UB00* - *No reason*
*UB01* - *Not codeable*
*PB20* - *Difference in means large (± 6 mm)*
*PB23* - *Means are not the same, even though std dev's (spread around means) are the same*
*PB26* - *Means not the same + ranges do not coincide (exact correspondence required)*
*PB27* - *Means not the same, even though there is some of overlap of their ranges*
*PB40* - *Possible values of two means (+/- up to 5mm) could be in same vicinity, or there could be a large difference, hence they don't agree*
*SB70* - *Intervals do not overlap in a large enough range*
*PB73* - *Means do not lie within range defined by other groups' intervals (mean ± σ)*

*Choice A: Point paradigm*

Students may have concluded that the two results agree based on the identical uncertainties (PA10). Others may have reasoned that the standard deviations imply that the variation in values are equal (PA13) and yet others that the large uncertainties indicate that both groups have not achieved the aim of determining the true value for *d* (PA14). This last response type is an extreme example of action and reasoning from the point paradigm; obtaining a deviation in results is seen as an experimental failure. The following responses were received from this cohort of students:

> *They both have the same standard deviation which implies the variation of their values is the same.* (PA13)

> *The two values agree in a sense that they are both wrong. It is obvious from the standard deviation of the values (it is not so small) that they are far from the true value of d.* (PA14)

Students may have viewed the difference in the means as small and indicated that the uncertainties are indeed identical (PA20). They could have argued that different apparatus or conditions caused the difference in the means since the standard deviations are the same (PA21). Respondents could have indicated that the minimum and maximum values for the means, by adding or subtracting 5 mm from the calculated means, yield a value close to the mean of the other group (PA23). Some may have observed that the means are close enough and that it is not necessary for the means to lie within the other groups' uncertainty intervals (PA27). The responses below are representative of this cohort:

> *The corresponding means are similar, but the standard deviations are exactly the same. This data agrees closely as there is very little difference between the point that the corresponding values are centralized.* (PA20)

> *The average may be different but they still got the same standard deviation of the average. Thus the difference in averages may be caused by using different apparatus.* (PA21)

> *By obtaining the maximum deviation for each mean value, it can be seen that the values are, in fact, similar in value:*
>
> $436 + 5\ mm = 441\ mm \approx 442\ mm$
>
> *And* $442 - 5\ mm = 437\ mm \approx 436\ mm$ (PA23)

*Their results are within the same order of magnitude at least. They are in fact very close and just because their intervals of uncertainty only just do not overlap, you cannot say that they disagree.* (PA27)

The written answers may have asserted that the "results" are similar or "nearly the same" without referring to the means or uncertainties (this in spite of the explicit instruction not to use the word "result" in the explanation) (PA60). This response was received:

*Yes, the results are nearly the same.* (PA60)

*Choice A: Set paradigm*

The students may have concluded that the values obtained by the groups agree to within experimental uncertainty (SA40). No explicit reference is made to the uncertainty intervals defined by the means and their standard deviations. A sample quote is shown below:

*To within experimental uncertainty, the values obtained for d by the different groups are the same.* (SA40)

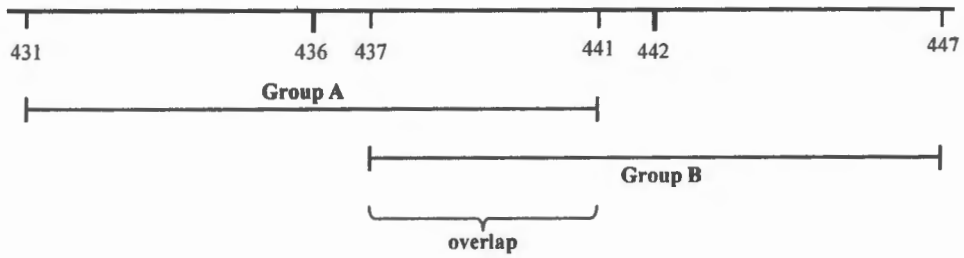
The conclusion made by students that the groups agree may have been based on the overlap of uncertainty intervals defined by the groups' means and standard deviations of the means (SA70). Respondents could further have suggested that the shift in results may be due to an under or over measurement by groups A and or B respectively (SA72). Others may have postulated that it is probable that the "true" value for *d* lies in the combined uncertainty range (SA74). The following quoted explanations are representative:

*There is an overlap of the two regions in which the groups have calculated that the means probably occur.* (SA70)

*Group A could have under measured and group B could have over measured but the std. deviation has an overlap between 437 and 441; this gives a 4 mm range of corresponding values, therefore they agree.* (SA72)

*The standard deviations of the means represent, roughly, the interval within which the true result could lie. Since these intervals overlap, the results do not contradict each other, and can be said to agree.* (SA74)

*The intervals of both answers overlap.* (SA70)




[Note: the diagram above is a nearly identical replica of the respondent's actual drawing in the space provided for explanation on the response sheet]

*Choice A: Undecided paradigm*

Siding with A does not categorise a response within a particular paradigm by default and hence the code UA00 would have been assigned for not furnishing an explanation, or coded UA01 for an unclear or non-categorisable response. None of the responses from this cohort fell into these categories.

*Choice B: Point paradigm*

The code PB20 was assigned when the difference in the means was considered as the main reason that the results of the measurement are incompatible. A few of these responses may have noted the equal standard deviations (PB23), required that the two intervals coincide (PB26) and some may have even mentioned the fact that the intervals overlap but stressed that this is immaterial in deciding compatibility (PB27). Sample quotes follow below:

*The mean values have ± 6 difference. This is a large difference.* (PB20)

*Although the std. deviations are similar, the means are quite different.* (PB23)

*The values for d differ and the range of answers do not overlap.* (PB26)
[Note: to this student "overlap" clearly means coincide]

*Although the range of values do overlap, they clearly don't agree. They are not the same value.* (PB27)

Some students may have compared the range of possible mean values defined by the uncertainties. These students thus considered some values in the ranges to be similar while others were viewed as being far apart; leading to a conclusion of incompatibility (PB40) as in the quotation below:

> *Even though the error may conclude that the means maybe is the same, it may also possibly conclude that they may be a great deal apart.* (PB40)

Respondents may have determined that one group's mean lies outside the other's interval of uncertainty, and or vice versa. Here, uncertainty intervals are acknowledged as part of the measurement result, but in essence the intervals are only used as a tool to judge the proximity of the means (PB73). The following couple of response quotes are representative:

> *Group B is correct in that the answers do not "agree" with each other. This is because both of the answers do not lie within the standard deviation of the other average.* (PB73)

> *436 + 5 = 441 < 442, likewise 442 – 5 = 437 > 436; therefore these two sets of readings are not close enough to be said to "agree". Ideally you would like both means to be within each other's standard deviations.* (PB73)

*Choice B: Set paradigm*

The action and underlying reasoning of students, who deem the degree of overlap of the two groups' uncertainty intervals as insufficient for compatibility, are clearly associated with the set paradigm, even though the conclusion is technically incorrect if the compatibility test is one standard uncertainty (SB70). The response below illustrates this:

> *More than half of the interval quoted by A lies outside of the range quoted by B and vice versa.* (SB70)

From the above, it is evident that responses to the DMSU probe could involve an action that follows from the set paradigm, but the reasons given may or may not have stemmed from that paradigm (see Figure 1.1). Buffler *et al.* (2001) describe this as an imposed set paradigm. In this work, however, as discussed in section 3.3.5 dealing with the SLG probe, due to the small sample size, paradigm codes primarily reflect underlying reasoning, hence the assignment of the point paradigm when students compared the two results using only the mean values, or only the numerical value of the standard deviations.

Table 3.9 groups the categories of ideas mainstream students used when considering the compatibility of two data sets presented in the formal manner after completing a full-year laboratory-based course which includes aspects of uncertainty.

*Table 3.9: Summary and frequency of code categories established for mainstream students' responses to the DMSU probe after instruction. (n =53)*

| Codes | Descriptor | Post-course |
|---|---|---|
| PA10-14 | Data sets are compatible if the standard deviations are the same. | 2 |
| PA20-27 | Data sets are compatible based on the perception that the difference in the means is "small" (various justifications). | 4 |
| PB20-27 | Data sets are incompatible based on the perception that the difference in the means is "large" (various justifications). | 8 |
| PB40 | Data sets not compatible since range of possible mean values not identical (some similar, others very different). | 1 |
| PB73 | Data sets are not compatible as the means do not lie in each other's uncertainty intervals. | 7 |
| SA40 | Compatibility based on the observation that the ranges of values of the two data sets are the same within experimental uncertainty. | 1 |
| SA70-74 | Data sets are compatible as the ranges defined by the means and uncertainties (standard deviation of mean) overlap in some region. | 28 |
| SB70 | Data sets are not compatible as the ranges defined by the means and standard deviations do not overlap at all. | 2 |
| | **Total** | **53** |

Inspecting the preceding frequency tables (Tables 3.1 to 3.9), it is evident that the individual responses to probes from this cohort of mainstream first year students was categorised within the point and set paradigm framework with very few that could not be so categorised. The consistency of use of paradigms across probes in the three areas of data collection, data processing and data set comparison required cross-probe analyses that involved assigning codes for a combination of probe responses.

## 3.4 Cross probe/overall paradigm assignment

### 3.4.1 Data Collection

The RT probe was only included in the pre-instruction questionnaire. Analysis of student responses showed that the placement of the RT probe in the order ($7^{th}$) might have resulted in students modifying their responses after seeing the data processing and data set comparison probes (UR, SMDS and DMSS), which presented the results of five repeated measurements, with the latter two including the calculated average or mean value. Some students, who consistently resorted to the point paradigm when answering the RD and RDA probes, used the set paradigm (repeat to calculate a mean value) when giving a response for the RT probe. A student responded to the RDA and RT probes as follows:

> *There is a discrepancy of 1 cm between the results, so another measurement is necessary to determine which one was more correct, so one has a reasonably accurate result in the end. The discrepancy is however not large enough to necessitate further releases.* (PB35d-RDA response)

> *The option of multiple measurements certainly applies once again to give a reasonably accurate average in the end, if accuracy counts a lot here for later calculations.* (SC25-RT response).

The pre-post comparison thus excluded the RT probe. However, previous studies have suggested that students may hold different epistemological views depending on the context of exercises (e.g. Leach *et al.*, 2000). Allie *et al.* (1998), in their study on special access students at UCT, found that significantly more students at entry to university saw the need to repeat time measurements in order to calculate a mean (set paradigm), when compared to the number that did so for distance measurements. Analysis of the RT probe thus only provided information on the nature and extent of differences, if any, in reasoning and actions used by mainstream students when performing different measurement-related activities (time and distance measurements).

Consequently, individual students' responses to only RD and RDA were considered together to establish paradigm usage in the area of data collection. Where responses to both probes were categorised as point (P), students were deemed to have used the point paradigm consistently. Similarly, consistent use of the set paradigm (S) was indicated for students that gave set responses to both RD and RDA. In certain instances, students used different paradigms when answering these two probes, indicating mixed paradigm usage (paradigm code M in the analysis). Inspection of

students' individual responses to the two data collection probes showed that this grouping of students (M paradigm code) made their procedural choices (number of repeats and the calculation of a mean value) based on a judgement of the relative proximity of the second measurement result presented in RDA. Most of these students responded to RD by either stating that repetition was not necessary as the same result was expected, or that one or more repeats were required only to confirm a recurring result (point paradigm). When confronted with the deviation in the second result presented in RDA, these students changed their minds and indicated the calculation of the mean value of several repeats (set paradigm). Interestingly, a small fraction of these students used the set paradigm to answer RD, but resorted to the point paradigm for RDA based on their contention that the deviation in the second result is too small too warrant further repetition and that the best result should be chosen from the previous two and one additional measurement. For the few cases where students gave responses to both probes that were not codeable within the point and set scheme, no conclusions could be made about overall paradigm usage (no definite paradigms - code X). This procedure made it possible to generate a table comparing the use of paradigms by individual students for data collection before and after instruction was generated to establish any shifts that may have occurred.

### 3.4.2 Data Processing

The rationale for paradigm code assignment in the area of data processing is similar to that used for the data collection probes; consistent point paradigm (P) for point responses to both the UR and SLG probes, consistent set paradigm (S) for set responses to both probes and mixed (M) for one point and one set response. Again, tables were generated comparing the use of paradigms for the data processing probes, before the course and after the course.

### 3.4.3 Data Comparison

Again, code assignment across probes followed a similar process as described before. Prior to instruction, responses of individual students to the SMDS and DMSS probes were considered together before assigning the combination code for data comparison. Students who recognised the spread in data as an indicator of the quality of a measurement result and used a measure of the spread in two data sets to gauge whether they were compatible, were classified as consistent set reasoners. Students who only considered individual data points or the means in their arguments for all the probes were classified as point reasoners. Mixed reasoning was indicated for students who recognised that the degree of dispersion in data sets reflects on the quality of a result, but failed to use overlapping intervals to determine whether two sets of data agreed with each other. Following

the code-assignment procedure, a table was generated to determine to what extent students' internalised the set paradigm for comparing data sets.

### 3.4.4 Paradigm usage across measurement phases

The latter phases of the analysis required that comparisons be made of paradigm usage in more than one area of measurement versus another. In these instances, not only the paradigm code, but also all the student's responses were considered to make a determination of overall paradigm usage. Consistent point, mixed and consistent set descriptors were used as described before. Finally, it was required to determine paradigm usage across all the probes, except the final probe (DMSU). This procedure will be described in Chapter 4, section 4.5.

# 4

# Findings

## 4.1 Overview

The previous chapter focussed on students' responses to individual probes; the framework of code assignment to the responses, the coding and categorization of responses and the rationale used to assign paradigm codes for clusters of probes within the three areas of data collection, data processing and data comparison. In this chapter the emphasis will be on comparing students' responses prior to instruction with those after instruction. In addition, other cross-probe analyses will be presented that illuminate mainstream students' knowledge state with regards to their understanding of measurement.

A key element of the analysis within the point/set framework is the grouping of probes within three areas of measurement: data collection, data processing and data set comparison. Section 3.4 presented the methodology used to assign paradigm codes for clusters of probe responses from individual students. In this way, patterns of paradigm use within the experimental aspects and across the entire set of probes were established. The objective was to determine the level of consistency of students' use of the point and set paradigms in making decisions when performing an experiment. The presence or absence of specific elements of the paradigms before and after instruction were identified, which could instruct the course developer on the students' ability at entry and the strengths and shortcomings of the course in terms of the desired outcomes of the first year laboratory curriculum.

## 4.2 Data Collection

As mentioned in Section 3.4.1, the RT probe provided information on how mainstream students' responses to data collection probes depended on context. Considering the placement of the RT probe within the whole set, it should not be surprising that only one in eighteen students (5.7%) thought that several repeats were not required. This finding seems to mirror the finding of Allie *et al.* (1998) that more students see the need to repeat time than distance measurements. However, in that study, the RT probe was answered first, before the data collection probes (RD and RDA). Consequently, the procedural actions of the mainstream students when answering the RT probe for the current work were considered together with their written responses in order to test context dependence. Individual students' procedural action and written justification to the RT probe were compared to those given for the RD and RDA probes. Just under half (49.1%) the students responded to all the data collection probes identically in terms of both action and reasoning used. Inspection of this group of students' actual written answers showed robust placement within a particular paradigm, independent of context. A quarter (24.5%) of the cohort opted for more repeats for time than distance measurements and 15% of the total sample explicitly referred to the observed deviation in the distance measurements to support their choice of procedural action. However, most of the students in the latter group (5 in 8) already indicated multiple repeats for distance measurements (9.4% of cohort). Over a quarter (26.4%) of the sample explicitly stated factors related to timing effects (human reaction time and/or operation of the stopwatch) in support of their belief that multiple time readings should be taken. However, less than half of these students (11.3% of sample) indicated a different procedural action (more or fewer repeats) than that taken for the distance probes. This suggests that neither procedural action nor any particular stated reason alone indicated whether a student was influenced by context. Students may have received procedural hints from preceding probes and stated certain reasons merely in support of some deeper underlying thinking.

Hence, considering both procedural differences and reasoning differences together, the following scenario emerged; only 2 in 15 (13.2% of total sample) students conclusively demonstrated that context plays a significant and determining role procedurally when collecting data in the laboratory. The analysis of this probe confirmed the importance of correct placement of probes within the order of a questionnaire. The researcher firmly believes that the RT probe should have been placed either immediately before or after the RD probe, to ascertain conclusively whether a student is influenced by context. The present data did however support the previously published finding (Allie *et al.*, 1998) that context plays a significant role in practical tasks, however the extent of this was not conclusively established for this cohort of mainstream students.

Table 4.1 presents the frequencies of student responses for the data collection probes, RD and RDA, in terms of the use of a point or set paradigm before instruction, against the responses by the same students after instruction. The table illustrates the relative shifts in individual students' use of a paradigm for their actions and reasoning.

*Table 4.1: Students' use of paradigms when collecting data (RD and RDA probes). (n = 53)*

| | | **Paradigm after instruction** | | | |
|---|---|---|---|---|---|
| | | *Point paradigm* | *Set paradigm* | *No definite paradigm* | **Total** |
| **Paradigm before instruction** | *Consistent point paradigm* | 1 (1.9%) | 3 (5.7%) | 3 (5.7%) | 7 (13.2%) |
| | *Mixed paradigms* | 0 (0.0%) | 6 (11.3%) | 1 (1.9%) | 7 (13.2%) |
| | *Consistent set paradigm* | 0 (0.0%) | 32 (60.4%) | 5 (9.4%) | 37 (69.8%) |
| | *No definite paradigm* | 0 (0.0%) | 2 (3.8%) | 0 (0.0%) | 2 (3.8%) |
| | **Total** | 1 (1.9%) | 43 (81.1%) | 9 (17.0%) | **53 (100%)** |

The data in the Table 4.1 shows that the use of the set paradigm prior to instruction by this cohort of mainstream students was high before, and even higher, after the laboratory course. Just seven out of ten (69.8%) of the students at entry gave responses compatible with the set paradigm as per the framework presented in Chapter 1. These students typically indicated the determination of a mean value to best represent the data. Alternatively, some students argued that the necessity of calculating a mean depended on whether or not a spread in the data from several repeated measurements occurred. These ideas were described in detail in Chapter 3. Slightly more than one in eight students (13.2%) employed mixed paradigms and the same number of students (13.2% of sample) used the point paradigm consistently prior to instruction.

Table 4.1 shows that the responses to both data collection probes after instruction of a sizable proportion of students could not be categorised as either point or set (17.0% of sample), whereas before the course almost all the students could be so categorised (100% - (13.2% + 3.8%)). Inspection of these students' actual responses showed that despite strict and clear protocol instructions, many students after the course gave shortened and incomplete written explanations as compared to those given prior to instruction. The researcher's observation is that the post-

instruction test was conducted around the time of the end-of-year laboratory examination, which could have prompted many students to give shortened answers in order to complete the exercise in minimum time. All of these students mentioned the need to perform multiple repeats to improve the result or results (in term of accuracy, precision and/or reliability). Considering that more than half of the students who answered in this fashion after instruction used the set paradigm consistently before the course, it is highly probable that these students used notions stemming from the set paradigm when providing their incomplete answers to the post-instruction data collection probes. Inspection of the same individual students' responses to other probes (data processing and data comparison) showed that all the students (9 students, or 17.0% of cohort) could be re-categorised by inference as having used the set paradigm when collecting data post-instruction.

Considering the table and discussion above, the post-instruction scenario may be interpreted as follows. After instruction, the overwhelming majority of the students (43 + 9 = 52 students, or 98.1% of cohort) used the set paradigm consistently with less than 2% persisting in using the point paradigm consistently for collecting data.

The mainstream laboratory course thus seems to have been effective in shifting most students, who used the point or mixed paradigms prior to instruction, to using the set paradigm consistently when collecting data after the first year laboratory course. Amongst the students who entered the course using the point paradigm exclusively, more than 85% adopted the set paradigm for making procedural decisions during the data collection phase of practical work after the course. All of the students who before the course drew on different paradigms, based on deviation in measurement results (mixed), shifted to consistent use of the set paradigm after the course. The few students (3.8% of sample) whose responses could not be classified according to the point and set model (no definite paradigm) prior to instruction also shifted to adoption of the set paradigm after instruction.

Therefore, the vast majority of students who either consistently used the point paradigm or mixed paradigms at entry shifted to using the set paradigm consistently after instruction. This finding generally agrees with the findings reported by Buffler *et al.* (2001). In contrast, the use of the point paradigm in that group of special access (GEPS) students, compared to mainstream students in the current study, was more prevalent both prior to and after a first year laboratory curriculum as evidenced by this statement, "Whereas before teaching, more than half of the students consistently used the point paradigm, only one in five did so after instruction" (Buffler *et al.*, 2001; p1143). Buffler *et al.* also found that for this cohort of special access students the largest percentage shift to the set paradigm after instruction occurred among the group who used the point paradigm consistently prior to teaching. They reported, "Deeper analysis of the data shows that the largest

shift from the use of a point to set paradigm occurred in the group of students who initially repeated to find the recurring value, but later chose to calculate a mean" (Buffler *et al.*, 2001; p1143). For the mainstream students in this study, although similar in terms of shift towards the set paradigm, a larger proportion of students at entry used different paradigms based on expectation of deviation in results. This is not surprising, as the mainstream students may have learned at their laboratory-equipped schools that deviation in results may be dealt with by calculating an average value. It is evident that in both cohorts the laboratory curriculum seems to have been effective in creating or strengthening in students' minds an appreciation for the spread inherent in data collection and the consequential need to represent data sets with a mean. The sample size of the mainstream cohort is relatively small; hence the remark above needs to be qualified by the realisation that the mainstream students may have entered the course with ad hoc routines for dealing with experimental results with the course merely providing further rule-of-thumb routines which students may have drawn on in answering the post-test probes. However, the level of reasoning used in mainstream students' responses to the data collection probes after the course compared to entry improved notably. This is evidenced by the fact that nearly a third of the students explicitly linked multiple repeats to the spread or standard deviation in the result/s after the course, whereas less than 4% did so at entry. These students recognised that gauging the spread (uncertainty) in a data set is an integral component of obtaining a measurement result.

The findings from a different study in the UCT-York project (Lubben *et al.*, 2001) showed that many students might have only adopted the set paradigm superficially. The calculation of a mean was perceived as an experimental requirement by many of the students in that cohort, and hence they repeated their measurements many times in order to satisfy the requirement of generating a mean. The same concern was raised with the responses from this cohort of students (see Section 3.3.1).

The analysis of the data processing and data set comparison probes presented in the following sections demonstrate to what extent the adoption of the set paradigm by this group of students was formulaic (i.e. rote learned) or whether the students acquired deeper levels of understanding in terms of the paradigmatic model.

## 4.3 Data Processing

Table 4.2 below contrasts the use of paradigm for answering the UR probe with that used for the SLG probe, by individual students at entry. The table suggests that the level of consistency of

mainstream students' use of paradigms across the data processing probes was less than that demonstrated for the data collection probes before instruction.

*Table 4.2: Students' use of paradigms when processing data prior to instruction. (n = 53)*

| | | **Paradigm used for UR** | | | |
|---|---|---|---|---|---|
| | | *Point paradigm* | *Set paradigm* | *Not codeable* | **Total** |
| **Paradigm used for SLG** | *Point paradigm* | 2 (3.8%) | 12 (22.6%) | 1 (1.9%) | 15 (28.3%) |
| | *Set paradigm* | 1 (1.9%) | 37 (69.8%) | 0 (0%) | 38 (71.7%) |
| | **Total** | 3 (5.7%) | 49 (92.5%) | 1 (1.9%) | **53 (100%)** |

Close to a quarter (22.6% + 1.9%) of the students at entry used mixed paradigms, the greater majority of them having used the set paradigm to answer the UR probe (calculation of a mean value) but resorted to the point paradigm when modelling a straight-line trend in a graphical set of data points (SLG) by forcing a line through points. This finding is not surprising, considering that the laboratory curriculum used by most South African schools do not emphasize the graphical modelling of data, but rather concentrates on dealing with multiple measurements by the calculation of mean values (92.5% of the students indicated a mean to represent a data set at entry). Consistent use of the set paradigm was high (69.8%) at the beginning of the laboratory curriculum however, with less than 4% of the students classified as consistently using the point paradigm for data processing tasks.

Table 4.3 displays the frequency of student responses for the data processing probes in terms of consistency of use of the point and set paradigms before instruction against responses by the same students after the laboratory course. The table sets out to demonstrate the level of proficiency that students exhibited in dealing with data sets and data representation and modelling before and after instruction.

*Table 4.3: Students' use of paradigms when processing data sets (UR and SLG probes). (n = 53)*

| | | **Paradigm after instruction** | | |
|---|---|---|---|---|
| | | *Mixed paradigms* | *Set paradigm* | **Total** |
| **Paradigm before instruction** | *Consistent point paradigm* | 2 (3.8%) | 1 (1.9%) | 3 (5.7%) |
| | *Mixed paradigms* | 3 (5.7%) | 10 (18.9%) | 13 (24.5%) |
| | *Consistent set paradigm* | 0 (0%) | 37 (69.8%) | 37 (69.8%) |
| | **Total** | 5 (9.4%) | 48 (90.6%) | **53 (100%)** |

The data presented in Table 4.3 shows that the use of the set paradigm increased substantially for both data processing probes after completion of the laboratory course (90.6% of students). None of the students used the point paradigm exclusively post-instruction. After instruction, over three quarters of the students (18.9% of sample) that used mixed paradigms at entry shifted to using the set paradigm throughout post-course. These students have clearly learned the straight-line fitting procedures introduced in their laboratory course. Only a few students (9.4% of cohort) persisted with the point paradigm for fitting a graphical trend in data points (SLG) but decided to calculate a mean value for a series of data readings (UR).

The post-course responses do however show that an overwhelming majority of mainstream students used the set paradigm in answering the data processing probes after completion of the laboratory course. Most of those who entered the course using elements of the point paradigm shifted to consistent use of the set paradigm post-instruction. This follows a similar trend as that observed for the shift in the paradigms used for data collection. By contrast, the responses of special access students as reported by Buffler *et al.* showed a larger proportion of students being consistent point data processors both prior to and after instruction.

## 4.4 Data set comparison

The probes studied prior to instruction include the SMDS and DMSS probes while after instruction the DMOS probe was added. Table 4.4 shows the frequency of students' responses classified according to use of paradigms before against that used by the same students after instruction. Only students who used the set paradigm consistently for answering *all the data set comparison probes* were classified as consistent set reasoners.

*Table 4.4: Students' use of paradigms when comparing data sets. (n = 53)*

| | | **Paradigm after instruction (SMDS + DMSS + DMOS)** | | | |
|---|---|---|---|---|---|
| | | *Consistent point paradigm* | *Mixed paradigms* | *Consistent set paradigm* | **Total** |
| **Paradigm before instruction (SMDS and DMSS)** | *Consistent point paradigm* | 8 (15.1%) | 6 (11.3%) | 4 (7.5%) | 18 (34.0%) |
| | *Mixed paradigms* | 1 (1.9%) | 28 (52.8%) | 6 (11.3%) | 35 (66.0%) |
| | *Consistent set paradigm* | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| | **Total** | 9 (17.0%) | 34 (64.2%) | 10 (18.9%) | **53 (100%)** |

Prior to instruction two-thirds of the students used mixed paradigms for data set comparison. Closer inspection revealed that almost all of these students recognised that the degree of spread in a data set is indicative of measurement quality as demonstrated in their responses to the SMDS probe, but failed to take into account all the data when deciding whether two data sets agreed for the DMSS probe. Students have at entry not yet received a course on data analysis so would lack the necessary knowledge to effectively compare data sets based on the degree of overlap of intervals. After the laboratory curriculum which incorporates a comprehensive data analysis course, students are expected to demonstrate an ability to assess the quality of data sets, to determine whether data sets agree or not, as well as to apply the formalistic tools of data analysis and interpretation. The expectation thus is that students would have used the set paradigm in all three of the probes (SMDS, DMSS and DMOS) after instruction.

The point paradigm was used throughout by a third (34% of sample) of the students and none were classified as consistent set reasoners when dealing with the quality and compatibility of data sets before instruction. A similar number (64% of sample) of students used mixed paradigms after instruction, 17% of students still persisted with point reasoning with only 1 in 5 students having adopted or internalised the set paradigm when comparing data sets.

More than 80% of the students that used mixed paradigms prior to instruction (1.9% + 52.8% of sample) still were unable to draw consistently on the set paradigm when reasoning about measurement across the three data set comparison probes after instruction. A slightly smaller proportion of the pre-course point reasoners (15.1% + 11.3% of sample) failed to come to terms with the higher conceptual demands of comparing data sets after completing their course. Inspection of students' actual responses revealed notably that more than half of the students who

displayed inconsistent use of the set paradigm after instruction explicitly referred to the formal constructs of standard deviation and/ or uncertainty in their written responses.

The data thus shows that the shifts observed for mainstream students having internalised a consistent set paradigm was small when compared to the shifts accomplished for data collection and data processing. The pre-post comparison of students' responses showed that only 22% (7.5 % of sample) of the point group and 17% (11.3% of sample) of the mixed group demonstrated a shift to the set paradigm across all the data comparison probes. The chi-squared statistic generated for these shifts ($\chi^2 = 0.200$) shows that the difference in the point and mixed group shifts was not statistically significant ($p = 0.654 >> 0.01$).

It thus appears that the course was not very successful in addressing the more deep-seated reasoning and understanding that underlies the decisions and actions required when considering the quality and compatibility experimental data.

## 4.5 Post instruction

It was reported in Sections 4.2 and 4.3 that the overwhelming majority of the mainstream students in this sample displayed a high level of competency in the areas of data collection and data processing after instruction as evidenced by the use of the set paradigm consistently across the RD and RDA probes in the one instance, and for the UR and SLG probes in the other. The question arose as to the level of consistency of the use of paradigms across these two areas of measurement. Lubben *et al.* (2001), in their study of a diverse group of university students at entry, reported that they found a strong link between the use of the point paradigm for data collection and data processing on the one hand, and the use of the set paradigm for both areas on the other ($\chi^2 = 70.5$, $p < 1 \times 10^{-15}$). Table 4.5 contrasts the mainstream students' use of paradigms for answering the data collection probes with that used by the same students when giving responses to the data processing probes, before the course.

*Table 4.5: Relationship between the use of paradigms in data collection and data processing before instruction. (n = 53)*

| | | **Paradigms used for data processing (from UR and SLG)** | | | |
|---|---|---|---|---|---|
| | | *Consistent point paradigm* | *Mixed paradigms* | *Consistent set paradigm* | **Total** |
| **Paradigms used for data collection (from RD and RDA)** | *Consistent point paradigm* | 3 (5.7%) | 0 (0%) | 4 (7.5%) | 7 (13.2%) |
| | *Mixed paradigms* | 0 (0%) | 0 (0%) | 7 (13.2%) | 7 (13.2%) |
| | *Consistent set paradigm* | 0 (0%) | 13 (24.5%) | 26 (49.1%) | 39 (73.6%) |
| | **Total** | 3 (5.7%) | 13 (24.5%) | 37 (69.8%) | **53 (100%)** |

A chi-squared test was performed on the data to determine whether the null hypothesis, that the use of paradigms in the two areas of measurement are not related at all, would be verified or rejected by the data at some level of significance (Muijs, 2004; Cohen *et al.*, 2000). The null hypothesis was rejected as the calculated $\chi^2$ value of 25.5 was greater than the critical value of 23.5 at the $p = 0.0001$ level of significance (i.e. $p << 0.01$) (Haslam and McGarty, 2003). A similar result was achieved for the post-course comparison between responses to the data collection and data processing probes (omitted to avoid repetition).

It was thus logical and easy to classify each student according to consistency of paradigm use over all the data collection and data processing probes used in the post instruction questionnaire. The previous section however indicates that this consistency of use of the set paradigm breaks down when data sets need to be qualified and compared. Table 4.6 presents the frequency of paradigms used across all the data collection and data processing probes; RD, RDA, UR and SLG, cross-tabulated against that used for the data set comparison probes, namely SMDS, DMSS and DMOS, after instruction. The table thus illustrates the relationship between the more formal aspects of measurement, i.e. data collection and processing; and the deeper understanding of measurement by way of data set comparison.

*Table 4.6: Students' use of paradigms for data collection / processing against those used for data comparison after instruction. (n = 53)*

| | | **Paradigms used for data collection and data processing (RD, RDA, UR and SLG)** | | | |
|---|---|---|---|---|---|
| | | *Consistent point paradigm* | *Mixed paradigms* | *Consistent set paradigm* | **Total** |
| **Paradigms used for data comparison (SMDS, DMSS and DMOS)** | *Consistent point paradigm* | 0 (0%) | 4 (7.5%) | 5 (9.4%) | 9 (17.0%) |
| | *Mixed paradigms* | 0 (0%) | 2 (3.8%) | 32 (60.4%) | 34 (64.1%) |
| | *Consistent set paradigm* | 0 (0%) | 0 (0%) | 10 (18.9%) | 10 (18.9%) |
| | **Total** | 0 (0%) | 6 (11.3%) | 47 (88.7%) | **53 (100%)** |

After instruction 88.7% of the sample consistently used the set paradigm for the data collection and data processing aspects/ requirements, i.e. the greater majority of students learned the formal tools of data analysis and appeared to know how to use them. All the students have at least adopted some elements of the set paradigm as evidenced by the fact that no student used the point paradigm consistently for data collection and processing after instruction. The remaining group (11.3 % of sample) represents those students that employed mixed paradigms for data collection and data processing post-course.

In contrast, nearly two-thirds of the students used mixed paradigms for the various situations posed by the data comparison probes, and only 18.9% of the students drew on the set paradigm consistently when comparing and interpreting data sets. The same students of the latter group (18.9% of sample) consistently based their responses on the set paradigm for all three areas of measurement studied, i.e. data collection, data processing and data set comparison. Nearly a sixth (17.0%) of the students persisted in using the point paradigm consistently for data set comparison, even after completing a course in basic data analysis.

The chi-squared analysis on the data in Table 4.6 ($\chi^2 = 12.1$, $p < 0.005$ for 2 degrees of freedom – the column of zeros omitted) suggests that the frequencies observed were statistically significant. A large proportion of students that used mixed paradigms for the data collection and data processing probes used the point paradigm consistently for answering the data comparison probes (two-thirds of mixed group or 7.5% of sample). None of the students in the mixed group could come to terms with the demands of data comparison. In contrast, a sizeable if small fraction of

students who applied the set paradigm consistently when collecting and processing data were also able to reason according to the set paradigm when comparing data sets (one fifth of the group or 18.9% of sample). More than two thirds of this grouping (60.4% of sample) employed mixed paradigms when comparing data sets. A smaller proportion of these students (11% of set group or 9.4% of sample) used the point paradigm consistently for data set comparison even though they used the set paradigm for data collection and data processing.

In summary, the data shows that only those students who came to terms with the more routine aspects of measurement (collecting and processing data) managed to grasp the more demanding concepts embedded in the data comparison probes. Stated differently, the chances of coping with tasks that requires a deeper, more fundamental understanding of measurement is very limited if students cannot acquire proficiency in the more routine aspects of experimentation.

Earlier, mention was made of the large proportion of students that used the formal language of data analysis to express their reasoning in terms of the spread in readings of a measurement. Yet, many of these students could not be considered to be grounded in the set paradigm in terms of how they used the formal constructs of standard deviation and uncertainty to make decisions on data set compatibility. Analysis of the DMSU probe, only included in the post-instruction questionnaire, explored how the students compared two sets of measurements that had been described in terms of a mean and a standard deviation. Table 4.7 contrasts individual students' use of a paradigm for answering DMSU compared to that used across all the previous probes (the two data collection probes, the two data processing probes and the three data set comparison probes described earlier).

*Table 4.7: Students' use of paradigms for data collection, data processing and data comparison after instruction. (n = 53)*

| | | **Paradigm used for DMSU probe** | | |
|---|---|---|---|---|
| | | *Point paradigm* | *Set paradigm* | **Total** |
| **Classification based on all previous probes (RD, RDA, UR, SLG, SMDS, DMSS and DMOS)** | *Inconsistent use of paradigms* | 21 (39.6%) | 22 (41.5%) | 43 (81.1%) |
| | *Consistent set paradigm* | 1 (1.9%) | 9 (17.0%) | 10 (18.9%) |
| | **Total** | 22 (41.5%) | 31 (58.5%) | **53 (100%)** |

It was shown before that it was relatively simple to ascertain a student's use of paradigms across all the data collection and data processing probes (RD, RDA, UR and SLG) due to the high level of consistency in use of paradigms across these probes. The use of paradigm for these two aspects of

measurement was combined with the paradigm usage across the three data set comparison probes (SMDS, DMSS and DMOS) to arrive at an overall representation of students' consistency of use of paradigms for all the probes. Based on previous tables, it is not surprising that none of the students were classified as consistent point reasoners after the laboratory course, 81.1% were inconsistent in their use of paradigms and slightly less than a fifth seem to have embraced the set paradigm convincingly. Even though 58.5% of the students used the set paradigm when being presented with data in the formal manner (mean +/- standard deviation of the mean), less than a third of this group (17.0% of sample) appear to be located firmly in the set paradigm. More than four-fifths of the students either used the point paradigm to answer the DMSU probe or correctly used the set paradigm by rote or in an *ad hoc* way (83% = 41.5% + 41.5%). This mirrors the finding of Buffler *et al.* (2001) where they reported that there seemed not to be any correlation between those students' ability to apply the formalistic rules of overlapping intervals and their underlying understanding of the statistical nature of measurement. The results from this cohort of mainstream students thus supports the observation made about the special access students (Buffler *et al.*, 2001), that an ability of students to reason appropriately when measurement results are presented in a formal way does not imply that the same students have developed a commensurate understanding of the underlying principles of their reasoning.

The last number and code qualifiers of the codes as explained in Chapter 3, provided information on the students' secondary reasons when justifying their procedural actions in the various phases of laboratory activities. Spreadsheets that contain the codes for the students' pre- and post-course responses to all the probes are provided in Appendix II. A quick inspection of the codes shows that more than 80% of the students before instruction, and over 70% after, used some notion of the term "accuracy" in support of some or all their arguments. Some students used the terms "accuracy" and "precision" to underscore the same reason, indicating that these terms are interchangeable in the students' minds (mirrors finding of Evangelinos *et al.*, 2002). Some students linked accuracy to the uncertainty interval evident in the spread in data, while others expressly implied that the spread is related to "errors", external factors beyond their control or "inaccuracies". It is interesting to note the greater number of students after instruction that explained their procedural actions on the basis of wanting to get closer to the "true value". There appears to be no link between the use of paradigms and the usage of technical terminology. In fact, the actual responses revealed that many students seemed to appeal to these terms only to give credence to their actions and reasoning, not demonstrating that they in fact understood the scientific meaning of using these terms. What must be of concern is that after the course this haphazard language usage seemed to have increased amongst the students; the course seemed to have entrenched the idea that experiments concern the attainment of results that are as close to the ideal as possible.

# 5

# Discussion

## 5.1 Mainstream students' views of measurement in terms of point and set paradigms

The first research question of this study explores the usefulness of the model of point and set paradigms for interpreting the ideas about measurement held by mainstream students. The data show that the majority of responses provided by the mainstream students to the probes were identifiable as being associated with either the point or set paradigms. Furthermore, almost all the written responses were both clear and detailed with very few ambiguous answers. The students generally displayed good writing skills, which assisted with the coding of their responses.

Large percentages of individual responses, particularly to the RD, RDA and RT probes, could not be classified as point (P) or set (S), but could be inferred from cross-probe analysis. The point and set classification scheme facilitated cross-probe comparisons for each individual student. When looking at probes together in particular areas, it was relatively easy to place most students in either the point or the set mode of action and or reasoning. The mainstream students were consistent in their use of paradigms across the probes that deal with data collection and data processing, which allowed for a single paradigm code to be assigned for paradigm usage across both these areas for each student. Further then, the cross-probe analysis allowed for the comparison of paradigm usage in the more routine aspects of measurement, i.e. collecting and processing data, with that used for the more demanding area of data set comparison. After the laboratory course, there was greater alignment of responses according to the set paradigm for the areas of data collection and data processing with more sophisticated reasoning used. It is thus reasonable to conclude that the

paradigmatic framework was effective in accounting for the actual development of students' procedural and reasoning abilities. This study supports the goal of instruction model, as presented in Figure 1.1, that a shift to use of the set paradigm may reliably reflect the development of students' underlying understandings in the various areas of measurement.

This study thus suggests that mainstream students' understanding of measurement during the three phases of data collection (the reasons for repeating measurements), data processing (calculation or graph) and data set comparison (quality and compatibility) may be characterised in terms of the point or set paradigms. The framework provided a useful tool for investigating and interpreting the mainstream students' decision-making processes and actions during measurement related activities before and after instruction. The present work supports a recommendation made in a previous study (Allie *et al.*, 1998), that the point and set paradigmatic model should be used to inform the development of a laboratory curriculum for mainstream students that effectively addresses the underlying understanding required when conducting investigative activities in a scientific domain.

However, particular problems related to the following of protocol instructions were experienced. Despite careful administration, a number of students chose not to strictly adhere to the guidelines. Several students, despite clear instructions to insert completed probe sheets into the brown envelope, opted to only do so at the end after completing all the questions. Although the researcher walked around the venue to check on this, a number of students at the end were observed still not having followed this instruction. It is not clear whether these students actually changed the answers to previous probes based on subsequent ones. The post-course questionnaire also included a larger number of probes (15), which may have resulted in many students not giving each probe the appropriate level of consideration. This is evidenced by the many responses that referred to the answers of previous probes (e.g. "*As I said before, many repeats are required*" RDA response), despite clear and explicit instructions to answer each question in full even if students believed the answers to be identical to previous ones.

Although the design and order of the probes in the questionnaire were specifically aimed at minimising the 'learning effect' of earlier probes on later ones, it is clear from the responses that some students indeed responded to later probes based on the procedural directions hinted at in earlier probes. Many students based their arguments on the experimental reality presented in previous probes (e.g. the code qualifier 'd' was appended to response codes where students based their arguments on the fact that a deviation in the distance readings was observed). The point and set framework does not account for this learning effect in responses, especially from one area of measurement to another.

The probes emphasise repeated measurements of a single result at the expense of single readings. The probes therefore do not investigate explicitly how students view a single reading in terms of the information that it may provide about the measurand. Repeated readings, especially where only five are given as in the probes, may reflect trends in the data, and consequently result in students using reasoning deemed inappropriate according to the set paradigm (e.g. if a student perceives a trend in the data presented in the UR probe, the first or last measurement may be selected; reasoning coded as being consistent with the point paradigm).

Students' underlying reasoning and ideas were inferred by considering their chosen procedural directions together with their written responses. Usage of particular keywords (e.g. accurate and precise) was not considered as basis for identifying a type of reasoning due to previously identified problems students have in using technical terminology. This may have lead to a skewed interpretation of responses as the researcher's personal biases played a more significant role in assigning the code categories further exacerbated by the use of a pre-determined analysis scheme (point and set paradigms). Additionally, accrediting only two paradigms to students' underlying reasoning may have disregarded alternative paradigms. However, the researcher believes that these concerns were adequately addressed in the analysis process.

## 5.2 The development of mainstream students' understanding of measurement

The second and third research questions of this study survey the development of mainstream students' views of measurement, and any differences between these views and those of GEPS students. The answers to these research questions will be summarised and discussed together.

Mainstream students at entry to university exhibit high levels of proficiency when collecting and processing data as evidenced by the high percentages of students that used the set paradigm consistently prior to instruction. This is to be expected since most of these students have had good schooling backgrounds where they would have been exposed to laboratory work (Kaunda and Ball, 1998) and have learned basic skills in dealing with measurements (e.g. calculation of a mean). However, this cohort did exhibit some deficiencies at entry in terms of procedural actions related to data processing. For example, a significant number of students were unable to model a graphical trend in data. This finding, consistent with the findings of Buffler *et al.* (1998) for GEPS students,

may be explained by the fact that school science laboratories place emphasis on the calculation of a mean and largely neglect other methods of treating data. The mainstream laboratory course seems to have been effective in shifting most of the students that did not use the set paradigm consistently in the areas of data collection and data processing prior to instruction to consistently making decisions and reasoning according to the set paradigm after the course.

As set out in Chapter 1, the mainstream laboratory curriculum consists primarily of experiment-based exercises focused on the gathering and processing of data. The rest of the activities provide the formal tools and analytical algorithms (dealing with dispersion, least-squares fitting procedure, etc.), which the students need to learn. The emphasis is on the communication of experimental evidence; most of the laboratory exercises and the laboratory examinations require the result of an experiment to be stated with a mean and a standard deviation of the mean. The positive findings that relate to the students' abilities to successfully apply these data analytical routines are therefore not surprising. The mainstream students also displayed greater proficiency in the two more routine areas of measurement (collecting and processing data) than did the GEPS students of a previous study (Buffler *et al.*, 2001) both before and after instruction. The results from this study parallels the conclusions of Buffer *et al.* in their analysis that both laboratory courses (different in content but not underlying approach) were effective in creating or strengthening in students' minds an appreciation for the spread inherent in data and the consequential need to represent data sets with a mean and to model a graphical trend in data with statistics-based procedures.

Although the results showed that a large majority of the students displayed proficiency in collecting and processing data, they did not come to terms with the deeper level understanding required when comparing data sets. These differences concur with the work published by Gott and Duggan (1995) on large numbers of secondary school students in the United Kingdom. Students could learn successfully many of the procedural aspects of experiments (e.g. methods of collecting data, control of variables, setting up of tables of results, processing of data, etc.) but hit a threshold when required to interpret data. The majority of the mainstream students both at entry and after the course, recognised spread as indicative of the quality of a data set but failed to grasp that the spread in a set of data points provides information on the measurement result, formalised by way of a theoretical model, which should then be used to arrive at a conclusion about compatibility of data sets. The mainstream students in this study thus seem to have learned aspects of the set paradigm by rote, especially when rules-of-thumb may be successfully applied in dealing with data sets. This finding confirms the concern expressed by Lubben *et al.* (2001) in consideration of the responses of the GEPS students of that study. Both cohorts of students may thus be placed in the set action region of Figure 1.1 but their placement in terms of set reasoning is brought seriously into question.

The cross-probe analysis of mainstream students' responses unmasked most of them as point reasoners that use set actions by rote when dealing with routine procedures. Further, when presented with measurement results in the formal manner, calculated means and standard deviations of the means, more than half the students were able to apply the rule of overlapping intervals to compare the two data sets. The course therefore seems to have been only effective in teaching the rules of data analysis. The data shows that only one in six mainstream students' adopted the set paradigm across all the areas of measurement after instruction. This outcome is very similar to that reported by Buffler *et al.* (2001) in their study on GEPS students at UCT.

Mainstream students, after completing their laboratory course, continued with a haphazard use of terms such as 'accuracy', 'precision' and 'uncertainty' in their responses. As these students are mostly English first language users (in contrast with the GEPS students in the Buffler *et al.* study), this cannot be ascribed to misinterpretations related to language. Most students persisted with the argument that repeating measurement is required to reduce or limit 'random' error or to increase accuracy and/or precision, and that experiments are about reducing "errors" and honing in on a true or correct value. Séré *et al.* (1993) and Tomlinson *et al.* (2001) reported similar findings. The students' use of the terms 'error', 'mistake' or 'inaccuracy' to refer to an observed spread in data sets suggests that they completed their course with fundamental misunderstandings about the nature of scientific measurement. Very few of these students fully grasped the central role that uncertainty plays in reporting evidence in science, as evidenced by their reasoning when comparing data sets. The mainstream students evidently viewed the calculation of a mean as a way of dealing with experimental 'error', which may be a barrier to appreciating the inherent uncertainty present in all measured quantities. This is similar to the finding reported by Evangelinos *et al.* (1998).

The findings discussed above imply that most of the mainstream students who had learned the formal rules of data analysis and interpretation did not acquire a deep and fundamental understanding of measurement. Also, a large proportion of these students expressed themselves in their responses by using the formal terminology, but the majority of this group were unable to demonstrate that they understood the underlying principles involved in these statistical constructs. Apparently, learning the data analysis techniques did not assist these students in understanding the concepts that underlie these statistical procedures. These findings again are consistent with those of Buffler *et al.* (2001), Evangelinos *et al.* (1998) and Séré *et al.* (1993).

However, the data from this mainstream cohort did show that it was very unlikely for those students who did not master the more routine aspects of measurement, namely collecting and processing data, to learn aspects of measurement that requires a fundamental understanding of the

quality of experimental data, namely comparing and interpreting data. In contrast, at least some of the students that became proficient in using the formal tools of data analysis grasped the more qualitative aspects of measurement. However this does not imply understanding of the operational tools of data analysis. This finding needs to be considered when developing any new course on measurement, as the results discussed in the previous paragraphs suggest that traditional courses are wholly inadequate in providing the foundation blocks for a thorough understanding of the inherent statistical nature of measurement. The present data, together with the stated goals expressed by Leach (1999), lead to the conclusion that laboratory courses should target the teaching of measurement more explicitly, that activities should address underlying concepts more directly, and that the current traditional methods of introducing students to scientific measurement indeed may lead to students reverting to their deeply-held epistemological beliefs about the nature of evidence in science (Elby, 2001).

## 5.3 The teaching of measurement and uncertainty

The hope expressed by Leach (1999), that engaging in measurement-related activities may lead to better understanding of the nature of science and hence measurement, thus seems to be misplaced, at least for a traditional laboratory course. This disappointing result may be explained by the work on school children in the UK (Gott *et al.*, 1995) that postulated that unless concepts of evidence are specifically taught, students are unlikely to gain an understanding of scientific evidence. The laboratory course completed by the mainstream physics students at UCT did not explicitly address the students' underlying understanding of measurement and uncertainty. The work of Deardorff (2001) is consistent with the findings reported in other studies, including the present work, that traditional laboratory courses do not impart an understanding of measurement and uncertainty to students. As noted by him (p106) "students often ignore the uncertainty of a measurement when evaluating a result, and they use arbitrary criteria to decide if a result is acceptable".

Recently the results of several more empirical studies (Buffler *et al.*, 2001; Deardorff, 2001; Evangelinos *et al.*, 2002; and Masnick and Morris, 2002) brought into doubt the supposition that students gain understanding through performing experimental procedures only. These studies showed that after having completed a traditional introductory physics laboratory course, students could apply the mechanistic operations of data analysis (e.g. calculating means and standard deviations, fitting straight lines, etc.), but lacked an appreciation of the nature of scientific evidence, in particular the central role of measurement uncertainty.

Similar conclusions are drawn from the work of Deardorff. For instance, Lippmann (2003) reported that even after completing a reflective laboratory course less than half of the students used the notion of range for comparing sets of results. Masnick and Morris (2002), in their study including college students, found that they were more likely to use characteristics of the data set (variability, range size, outliers) for comparison and interpretation when they had little or no domain knowledge of measurement. This implies that students who enter university with little laboratory experience may be more likely to learn the underlying concepts of data set comparison and uncertainty than those that have pre-existent rules-of-thumb for data analysis. Students seem to reason intuitively about data when they lack knowledge and only learn how to apply formulaic rules after a standard course on data analysis. However, Masnick and Morris also found that novice students are unsure about compatibility of two overlapping data sets. Thus in order to remediate the lack of understanding of compatibility of data sets in mainstream students, concepts of measurement that deal with uncertainty and overlapping intervals need to be explicitly taught, as these seem to be counter-intuitive to students.

Recently, Case and Marshall (2004) reported on studies that investigated the approaches to learning of groups of second year South African chemical engineering students. Besides the traditional approaches of deep and surface learning, they identified two intermediary approaches that students use in problem solving contexts. Essentially deep learning involves understanding, whereas surface learning does not. Some of their students seemed to use either 'procedural surface' (algorithmic) or 'procedural deep' approaches when solving problems, depending on the course context. Both approaches have a focus on being able to solve problems, but the intention of the former is to remember formulae and solution methods, and the latter is to gain understanding through the application of solution methods. What is important is that the researchers found a strong link between the learning approaches adopted by students and the course context. An analogy may be drawn between problem solving and the taking of measurement decisions. The mainstream laboratory course could be considered as being structured around the teaching and assessment of routines, for which students need to demonstrate that they have learned the experimental procedures. The students respond by using a procedural surface learning approach, as the course does not necessarily require commensurate understanding. This suggests that the mainstream students have been disadvantaged by their course, as the activities they took part in did not support understanding of the nature of scientific evidence and uncertainty. What is more striking about the findings of Case and Marshall is that courses which have 'procedural deep' objectives may preclude some students from adopting a deep approach to learning, i.e. the learning and understanding of the underlying concepts. This suggests that practical courses that only focus on procedural activities, even if structured and targeted specifically at conceptual development,

may not stimulate all students to grapple with the more difficult and fundamental aspects of measurement.

Both the GEPS and mainstream laboratory courses (detailed in Sections 1.3 and 1.4) share the same underlying theoretical approach used in most university courses, the statistics-based, "frequentist" approach to measurement and uncertainty (discussed in next section). When the findings reported by Buffler *et al.* (2001) are compared to the results reported in this study, it clear that the mainstream students entered and exited their first year course more able to apply the formal procedures of data analysis than the GEPS students evidenced before and after their course. However, both groups were remarkably similar in their lack of understanding of how uncertainty relates to measurement results and what this implies for reporting of scientific evidence.

When considering recommendations based on the findings of this study, it is appropriate to again consider the laboratory course that the mainstream students followed, in contrast with the course the GEPS students completed. The GEPS laboratory course, applicable for evaluating the findings reported in the Buffler *et al.* (2001) study, addressed the more procedural aspects of measurement directly, by way of specifically targeted exercises and laboratory tasks which were designed to introduce the students to the rudiments of performing measurements and to guide these students in learning the basic elements of experimentation. The specificity of the tasks and the foregrounding of the skills and procedures involved in scientific measurement thus satisfied the recommendations made by several research groups over the last decade concerning the orientation of practical courses around clearly defined purposes (Osborne, 1996; Hodson 1998). The GEPS laboratory course also attempted to explicitly link the students' development of practical skills with their understanding of evidence as the UCT-York research group believes that experimental skills comprise a distinct body of knowledge that need to be explicitly taught, as also argued by Gott and Duggan (1996) and others. The mainstream course on the other hand, assuming higher levels of prior laboratory experience by these university entrants, is structured around experiment-based tasks aligned closely with the theory component of the course. The mainstream students thus completed a traditional introductory physics laboratory curriculum. The only aspects of measurement explicitly taught in the mainstream physics course are the statistical algorithms and data analytical tools that the students would require in reporting the results of their experimental tasks. Pedagogically, the mainstream course then assumes that students will develop the necessary skills and hence understanding related to scientific evidence, by 'doing practical work'.

As reported in this dissertation, most of the mainstream students entered university with previously learned methods of dealing with data (rote set actions, see Figure 1.1). The fact that so few have

fully internalised the set paradigm for making decisions in the laboratory, may suggest that the 'rules of thumb' these student have acquired at school, together with those learned during their laboratory course, may have seriously impeded the students' progress in acquiring a sound grounding in experimental measurement at university. These techniques of dealing with experimental measurements seem to be entrenched in students' minds and suggest that it is grounded in an epistemological view of measurement. Consequently, it is not surprising that the traditional physics laboratory course at UCT was unable to shift most of these students to coherent use of the set paradigm. Again, these observations are consistent with the findings reported by Hammer and Elby (2003), Ryder and Leach (2000) and Leach *et al.* (2000). Students generally tend to ignore the central role of theoretical models in interpreting data.

## 5.4 A probabilistic approach to teaching measurement and uncertainty

The UCT-York research group, by probing diverse groups of students' understanding of measurement, framed the constructs of point and set paradigms to account broadly for the two views of measurement students may hold. The elements of these two paradigms were discussed in some detail in Section 1.2. Fundamentally, the two paradigms differ in that with the point paradigm conclusions are drawn from individual data points or values, while with the set paradigm, properties constructed from the whole ensemble of data are used to inform knowledge of the measurand. An important feature of the set paradigm is the construction of a *distribution* from the data, from which the best approximation of the measurand and an interval of uncertainty are derived.

### 5.4.1 The "frequentist" approach to measurement

With the traditional or "frequentist" approach, a large number of repeated measurements are required to generate a frequency distribution that approximates that of the traditional theoretical model. The data analysis is typically based on the classical Gaussian distribution, which assumes an infinite number of readings. However, in nearly all practical situations in the introductory laboratory, the best approximation of the measurand will either be the reading itself (in the case of a single reading) or the calculated average value of a set of repeated readings, and then too few in number (typically not more than 5 or 10) to apply the statistical model with confidence. There is no coherent method provided for students to construct a meaningful distribution from the results of their actual experiments and they cannot model the result of a single measurement using frequentist

data analysis. Students' actual practice in the laboratory is thus in conflict with the theory upon which they are expected to base their data analysis and interpretation. The logical inconsistencies in the traditional approach to data treatment may thus further cultivate students' misconceptions about measurement in the scientific context.

### 5.4.2 The "probabilistic" approach to measurement

Whereas frequentist data analysis is based on frequency distributions, the "probabilistic" approach to measurement as the term implies is based on probability distributions. One of the attractive features is that the theory based on probability distributions is applied in the same way to a single reading as to an ensemble (Allie *et al.*, 2003). In 1993, the International Organization for Standardization (ISO) published recommendations for reporting measurements and uncertainties based on the *probabilistic interpretation of measurement*, in response to the need for a consistent international language for evaluating and communicating measurement results. All international standards bodies, including the IUPAP (International Union of Pure and Applied Physics) and IUPAC (International Union of Pure and Applied Chemistry), have subsequently adopted these recommendations for reporting scientific measurements. A number of documents currently serve as international reference standards. The most widely known are the so-called International Vocabulary of Basic and General Terms in Metrology (ISO, 1993) and the Guide to the Expression of Uncertainty in Measurement (ISO, 1995). A shorter version of the latter is publicly available as NIST Technical Note 1297 (Taylor and Kuyatt, 1994).

The recommended approach (ISO 1993, 1995) to metrology is based on the use of probability theory and the concept of the probability density function for the analysis and interpretation of data. A key element of the ISO guidelines is how it views the measurement process. The guidelines state, "In general, the result of a measurement is only an approximation or estimate of the value of the specific quantity subject to measurement, that is, the measurand, and thus the result is complete only when accompanied by a quantitative statement of its uncertainty." Uncertainty is defined as, "a parameter associated with a measurement result, that characterizes the dispersion of the values that could reasonably be attributed to the measurand" (ISO 1993, 1995). At the beginning of the measurement process, new data are combined with all prior information about the measurand to form an updated state of knowledge from which inferences about the measurand are made (see Figure 5.1). The formal mathematics used to allow these inferences are probability density functions (pdf's) with the (true) value of the measurand as the independent variable (in the ISO guidelines, the terms, "the value of the measurand" and "the true value of the measurand", have the same meaning). Thus, the measurement process includes using a pdf, which best represents our

knowledge about the measurand. Both the case of the single reading and the case of a set of repeated readings with dispersion, involve seeking the pdf for the measurand. The last step in the measurement process involves making inferences about the measurand based on the (final) pdf.

Although the ISO recommendations do not refer explicitly to the underlying philosophy, the formalism relies on the Bayesian approach to data analysis (see for example d'Agostini, 1999). The final pdf is usually characterized in terms of its location, an interval along which the (true) value of the measurand may lie, and the probability that the value of the measurand lies on that interval. In metrological terms these are, respectively, the best estimate (or best approximation) of the measurand and its uncertainty, and the coverage probability (or level of confidence), calculated as the percentage area under the pdf defined by the uncertainty interval. Typical statements describing a measurement result are of the form "the best estimate of the value of the measurand is $X$ with a standard uncertainty $U$ and the probability that the measurand lies on the interval $X \pm U$ is $Z$ %". In this approach, instrument readings are considered as constants, while the concept of probability is applied to any claims made about the value of the measurand, which is considered the random variable. With the probabilistic treatment of data analysis, there is no conflict between "exact" readings and uncertainty in measurement results. Uncertainty is related to knowledge of the measurand, not the reliability of the data. The term error is removed from the treatment of data and replaced with the more neutral term uncertainty. Further, the confusion between concepts such as precision and accuracy, and systematic and random error, are avoided.

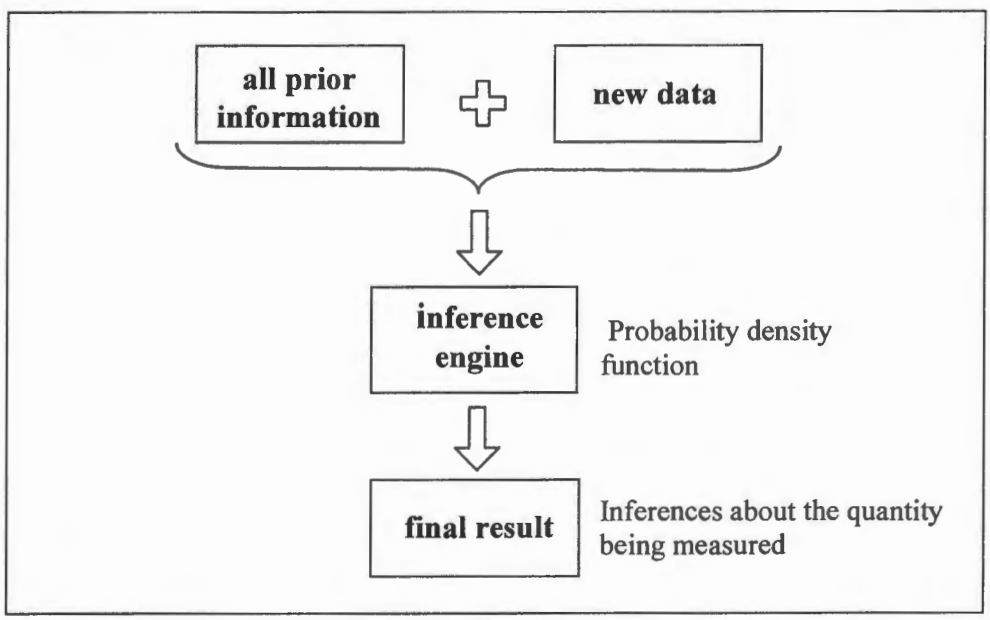



*Figure 5.1: A model for determining the result of a measurement (adapted from Allie et al., 2003).*

### 5.4.3 Evaluation of courses teaching the probabilistic approach to measurement

In response to the mismatch between the desired outcome of students gaining a thorough understanding of measurement and the findings that students lack such understanding, a few research groups, principally the research group at the University of Thessaloniki in Greece as well as the UCT-York group, have attempted to develop new laboratory curricula, based on the probabilistic interpretation of data. Since the ISO have adopted the probabilistic approach for reporting the results of scientific investigations, it was deemed doubly appropriate to base any new course on measurement on the guidelines set down by international standards bodies. Deardorff (2001) reported that most of the physics instructors (and students) surveyed in his samples were unfamiliar with the ISO recommended practices (based on the probabilistic interpretation of measurement).

Evangelinos *et al.* (2002) reported that students demonstrated higher levels of appreciation for the role of uncertainty in measurements after completing a probabilistic course in measurement. In contrast, students who followed the conventional course in their study ignored the probabilistic nature of a measurement result and rather viewed data analysis as a formal means of expressing the extent to which the measured value approximates the 'true value', or to achieve the true value itself. These students persisted with the notion of 'error' being a measure of how well an experiment was performed (the size of the spread), and that the formal tools of data analysis primarily provide the basis for reporting this. With their innovative laboratory course, the teaching of the term 'error' was discarded in favour of the metrological concept of uncertainty and the foregrounding of the concept of probability avoided the connotation between uncertainty and the spread observed in repeated measurement results.

The UCT-York group also developed a probabilistic course in measurement (Buffler *et al.* 2005). Buffler *et al.* (2003) again used the point and set paradigmatic model to assess the extent to which GEPS students adopted the set paradigm in all the areas of measurement after completing the probabilistic course. The results of this study showed that considerably more students (89%) could be categorised as consistent set reasoners after the probabilistic course in measurement, compared to that for the traditional course (16%; reported in Buffler *et al.*, 2001). However, the researchers caution that their data did not establish whether students who used the concept of overlapping intervals correctly have in fact internalised the set paradigm or used it by rote. They reported that inspection of the students' responses to other probes showed that the students might indeed have had an inappropriate understanding of an interval. However, the students did demonstrate that they

have learned the metrological aspects of measurement reporting as set down by the international standards bodies.

## 5.5 Conclusion

In conclusion, this study supports the findings published by the UCT-York group (e.g. Allie *et al.*, 1998; Buffler *et al.*, 2001, Lubben *et al.*, 2001) and other independent studies (Davidowitz *et al.*, 2001; and Rollnick *et al*, 2001) that the point and set model is a reliable and useful tool for classifying students' actions and reasoning when conducting measurement activities in the physics as well as chemistry laboratories. Redish (2003) further reports that the set of probes developed by the group together with the coding schemes, provides a validated diagnostic tool for investigating students' ideas of measurement.

The fact that the point and set paradigms form the basis of students' thinking in both physics and chemistry contexts supports the view that measurement relies on a specific domain of knowledge, rather than subject-specific skills (see Gott and Duggan, 1996). The current study on mainstream physics students support previous studies (e.g. Buffler *et al*, 2001; and Buffler *et al.*, 2003) in stating that the main aim of introductory physics courses should be to shift students' use of the point paradigm to that of a set paradigm when making measurement decisions.

The traditional, frequentist type course was not very successful in enabling such a shift, despite the mainstream students' higher levels of preparedness and understanding at entry to university. The current laboratory course for mainstream physics students at UCT, cannot provide the conceptual framework for the development of student understanding of measurement and a deeper appreciation for the inherent statistical nature of experimental evidence, since it does not address concepts of evidence directly, and encourages students to create *ad hoc* routines to deal with measurement. With no fundamental change in epistemological beliefs about the nature of measurement, as the results of this study suggest, the mainstream students may have reverted to ingrained learning methodologies in more advanced courses (Elby, 2001). The probabilistic course developed, piloted and validated with GEPS students (Buffler *et al.*, 2003) at UCT may provide a better platform for establishing in these students' minds a sound understanding of measurement. If the results of the present work are a true reflection of the students' abilities at the end of their first year of study, then the course may have in fact limited the students' progress in learning physics and other science courses, which require a solid grounding in experimentation.

# References


Allie, S. and Buffler, A. (1998) A course in tools and procedures for Physics 1. *American Journal of Physics*, 66 (7), 613-624.

Allie, S., Buffler, A., Campbell, B., Lubben, F., Evangelinos, D., Psillos, D. and Valassiades, O. (2003) Teaching measurement in the introductory physics laboratory. *The Physics Teacher*, 41 (7), 394-401.

Allie, S., Buffler, A., Kaunda, L., Campbell, B. and Lubben, F. (1998) First-year physics students' perceptions of the quality of experimental measurements. *International Journal of Science Education*, 20 (4), 447-459.

Allie, S., Buffler, A., Kaunda, L. and Inglis, M. (1997) Writing-intensive physics laboratory reports: tasks and assessment. *The Physics Teacher*, 35 (7), 399-405.

Allie, S., Buffler, A., Lubben, F. and Campbell, B. (2001) Point and set paradigms in students' handling of experimental measurement. In H. Behrendt, H. Dahncke, R. Duit, W. Graber, M. Komorek, A. Kross and P. Reiska (Eds.) *Research in Science Education: Past, Present and Future*. Kluwer Academic Publishers: Dordrecht. pp. 331-336.

APU (1988) *Science at age 13: review report.* London: Department of Education and Science.

Arons, A. B. (1993) Guiding insight and inquiry in the introductory physics laboratory. *The Physics Teacher*, 31 (7), 278-282.

Bell, J. (1999) *Doing your own research project: a guide for first time researchers in education and social science* (3rd edition). Buckingham: Open University Press.

Black, P. (1993) The purposes of science education. In R. Hull (Ed.), *ASE Secondary Science Teachers' Handbook.* London: Simon & Schuster.

Buffler, A., Allie, S., Campbell, B. and Lubben, F. (1998) The role of laboratory experience at school on the procedural understanding of pre-first year science students at UCT. In N.A. Ogude and C. Bohlmann (Eds.): *Proceedings of the 6th Annual Meeting of the Southern African Association for Research in Mathematics and Science Education*. pp 495-502.

Buffler, A., Allie, S., Lubben, F. and Campbell, B. (2001) The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, 23 (11), 1137-1156.

Buffler, A., Allie, S., Lubben, F. and Campbell, B. (2003) *Evaluation of a research-based curriculum for teaching measurement in the first year physics laboratory.* Paper presented at the 4th Conference of the European Science Education Research Association, Noordwijkerhout, The Netherlands, August 2003.

Buffler, A., Allie, S., Lubben, F. and Campbell, B. (2005) *Introduction to Measurement in the Physics Laboratory: A Probabilistic Approach.* Department of Physics, University of Cape Town. Available from: http://www.phy.uct.ac.za/people/buffler/labmanual.html

Campbell, B., Allie, S., Buffler, A., Kaunda, L. and Lubben, F. (2000) The communication of laboratory investigations by university entrants. *Journal of Research in Science Teaching*, 37 (8), 839-853.

Case, J. and Marshall, D. (2004) Between deep and surface: procedural approaches to learning in engineering education contexts. *Studies in Higher Education*, 29 (5), 605-615.

Chi, M., Feltovitch, P. and Glaser, R. (1981) Categorization and representation of physics problems by experts and novices. *Cognitive Science,* 5, 121-152.

Coelho, S. and Séré, M-G. (1998) Pupils' reasoning and practice during hands-on activities in the measurement phase. *Research in Science and Technological Education*, 16 (1), 79-96.

Cohen, L., Manion, L. and Morrison, K. (2000) *Research methods in education*. 5th edition. London: Routledge.

d'Agostini, G. (1999) *Bayesian Reasoning in High Energy Physics – Principles and Applications.* CERN Yellow Report 99-3: Geneva.

Davidowitz, B., Lubben, F. and Rollnick, M. (2001) Undergraduate science and engineering students' understanding of the reliability of chemical data. *Journal of Chemical Education*, 78 (2), 247-252.

Deardorff, D. L. (2001) *Introductory Physics Students' Treatment of Measurement Uncertainty.* Unpublished PhD thesis: North Carolina State University.

Elby, A. (2001) Helping physics students learn how to learn. *American Journal of Physics*, 69 (7), S54-S64.

Etkina, E., Van Heuvelen, A., Brookes, D. and Mills, D. (2002) Role of experiments in physics instruction - a process approach. *The Physics Teacher*, 40, 351-355.

Evangelinos, D., Psillos, D. and Valassiades, O. (1998) Students' introduction to measurement concepts: A metrological approach. In European Commission Report on Project PL 95-2005 *Labwork in Science Education*. pp. 561-587.

Evangelinos, D., Psillos, E. and Valassiades, O. (2002) An investigation of teaching and learning about measurement data and their treatment in the introductory physics laboratory. In D. Psillos and H. Niederrer (Eds.): *Teaching and learning in the science laboratory.* Dordrecht: Kluwer Academic Publishers. pp. 179-190.

Fairbrother, R. and Hackling, M. (1997) Is this the right answer? *International Journal for Science Education*, 19 (8), 887-894.

Garrett, J., Horn, A. and Tomlinson, J. (2000) Misconceptions about error. *University Chemistry Education,* 4 (2), 54-57.

Germann, P. and Aram, R. (1996) Student performances on the science processes of recording data, analysing data, drawing conclusions and providing evidence. *Journal of Research in Science Teaching*, 33 (7), 773-798.

Germann, P., Aram, R. and Burke, G. (1996) Identifying patterns and relationships among the responses of seventh-grade students to the science process skill of designing experiments. *Journal of Research in Science Teaching*, 33 (1), 79-99.

Gerson, R. and Primrose, R. A. (1977) Results of a remedial laboratory program based on a Piaget model for engineering and science freshman. *American Journal of Physics*, 45 (7), 649-651.

Gillham, B. (2000) *Developing a Questionnaire*. London: Continuum.

Gott, R. and Duggan, S. (1995). *Investigative work in the science curriculum.* Buckingham: Open University Press.

Gott, R. and Duggan, S. (1996). Practical work: its role in the understanding of evidence in science. *International Journal of Science Education*, 18 (7), 791-806.

Gott, R., Duggan, S. and Johnson, P. (1999) What do practicing applied scientists do and what are the implications for science education? *Research in Science and Technology Education*, 17 (1), 97-107.

Gott, R., Duggan, S., Miller, R. and Lubben, F. (1995) Progression in investigative work in science. In P. Murphy, M. Sellinger, J. Bourne and M. Briggs (Eds.): *Subject Learning in the Primary Curriculum: Issues in English, Science and Mathematics*. London: Routledge.

Hammer, D. (1994) Epistemological beliefs in introductory physics. *Cognition and Instruction,* 12, 151-183.

Hammer, D. and Elby, A. (2003) Tapping epistemological resources for learning science. *Journal of the Learning Sciences*, 12 (1), 53-90.

Haslam, S. A. and McGarty, C. (2003) *Research Methods and Statistics in Psychology*. London: Sage Publications.

Heller, P., Keith, R. and Anderson, S. (1992) Teaching problem solving through cooperative grouping. 1. Group versus individual problem solving. *American Journal of Physics*, 60 (7), 627-636.

Hestenes, D. (1987) Toward a modeling theory of physics instruction. *American Journal of Physics*, 55 (5), 440-454.

Hodson, D. (1998) Taking practical work beyond the laboratory. *International Journal of Science Education,* 20 (6), 629-632.

International Organization for Standardization. (1993) *International Vocabulary of Basic and General Terms in Metrology (VIM)*. ISO: Geneva.

International Organization for Standardization (1995) *Guide to the expression of uncertainty in measurement (GUM)*. ISO: Geneva.

Kaunda, L. and Ball, D. (1998) An investigation of students' prior experience with laboratory practicals and report writing. *South African Journal of Higher Education*, 12, 130-139.

Kaunda, L., Allie, S., Buffler, A., Campbell, B. and Lubben, F. (1998) An investigation of students' ability to communicate science investigations. *South African Journal of Higher Education*, 12 (1), 122-129.

Kirschner, P. and Huisman, W. (1998) Dry laboratories in science education: computer-based practical work. *International Journal of Science Education*, 20 (6), 665-682.

Kuhn, D., Amsel, E. and O'Loughlin, M. (1998) *The Development of Scientific Thinking Skills*. London: Academic Press.

Larkin, J. and Reif, F. (1979) Understanding and teaching problem solving in physics, *European Journal of Science Education*, 1, 191-203.

Laws, P. (1996) Millikan Lecture 1996: Promoting active learning based on physics education research in introductory physics courses. *American Journal of Physics*, 65 (1), 14-21.

Leach, J. (1999) Students' understanding of the co-ordination of theory and evidence in science. *International Journal of Science Education*, 21 (8), 789-806.

Leach, J., Millar, R., Ryder, J. and Séré, M-G. (2000) Epistemological understanding in science learning: the consistency of representations across contexts. *Learning and Instruction*, 10, 497-527.

Lippmann, R. (2003) *Students' understanding of measurement and uncertainty in the physics laboratory: social construction, underlying concepts, and quantitative analysis.* Unpublished PhD thesis: University of Maryland.

Long, D. D., McLaughlin, G. W. and Bloom, A. M. (1986) The influence of physics laboratories on student performance in a lecture course. *American Journal of Physics*, 54 (2), 122-125.

Lubben, F. and Millar, R. (1996) Children's ideas about the reliability of experimental data. *International Journal of Science Education*, 18, 955-968.

Lubben, F., Campbell B., Buffler, A. and Allie, S. (2001) Point and set reasoning in practical science measurement by entering university freshman. *Science Education*, 85 (4), 311-327.

Masnick, A. and Morris, B. (2002) Reasoning from data: the effect of sample size and variability on children's and adults' conclusions. In W. Gray and C. Schunn (Eds.): *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum. pp. 643-648.

McDermott, L. C. (1991) What we teach and what is learned – Closing the gap. *American Journal of Physics*, 59 (4), 301-315.

McDermott, L. C. and Redish, E. F. (1999) Resource Letter: PER-1: Physics Education Research. *American Journal of Physics*, 67 (9), 755-767.

Meester, M. and Maskill, R. (1995) First year chemistry practicals at universities in England and Wales - aims and the scientific level of the experiments. *International Journal of Science Education*, 17 (5), 575-588.

Millar, R., Gott, R., Lubben, F. and Duggan, S. (1996) Children's performance of investigative tasks in science: A framework for considering progression. In M. Hughes (Ed.), *Progression in Learning*. Clevedon: Multilingual Matters. pp. 82-108.

Millar, R., Le Marechal, J-F. and Tiberghien, A. (1999) 'Mapping' the domain: varieties of practical work. In D. Psillos and H. Niederrer (Eds.): *Teaching and learning in the science laboratory*. Dordrecht: Kluwer Academic Publishers. pp. 33-59.

Millar, R., Lubben, F., Gott, R. and Duggan, S. (1994) Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. *Research Papers in Education*, 9 (2), 207-248.

Muijs, D. (2004) *Doing Qualitative Research in Education*. London: Sage Publications.

Oppenheim, A. N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. London: Pinter Publishers.

Osborne, J. (1996) Untying the Gordian knot: diminishing the role of practical work. *Physics Education*, 31 (5), 271-278.

Patton, M. W. (1980) *Qualitative Evaluation Methods*. New York: Sage Publications.

Pfundt, H. and Duit, R. (1994) *Bibliography: Students' Alternative Frameworks and Science Education*. Kiel: IPN.

Redish, E. (2003) *Teaching physics with the physics suite*. Hoboken, NJ: Wiley.

Reif, F. and Larkin, J. H. (1991) Cognition in scientific and everyday domains: comparison and learning implications. *Journal of Research in Science Teaching*, 28 (9), 733-760.

Reif, F. and St. John, M. (1979) Teaching physicists' thinking skills in the laboratory. *American Journal of Physics*, 47 (11), 950-957.

Rollnick, M., Dlamini, B., Lotz, S. and Lubben, F. (2001) Views of South African chemistry students in university bridging programmes on the reliability of experimental data. *Research in Science Education*, 31 (4), 553-573.

Roth, W. and Roychoudhury, A. (1993) The development of science process skills in authentic contexts. *Journal of Research in Science Teaching*, 30, 127-152.

Roth, W-M., McRobbie, C. J., Lucas, K. and Boutonné, S. (1997) Why may students fail to learn from demonstrations? *Journal of Research in Science Teaching*, 34 (5), 509-533.

Ryder, J. and Leach, J. (2000) Interpreting experimental data: the views of upper secondary school and university science students. *International Journal of Science Education*, 22 (10), 1069-1084.

Séré, M-G., Fernandez-Gonzalez, F., Gallegos, J., Gonzalez-Garcia, F., De Manuel, E., Perales, J. and Leach, J. (2001) Images of science linked to labwork: a survey of secondary school and university students, *Research in Science Education*, 31, 499-523.

Séré, M-G., Journeaux, R. and Larcher, C. (1993) Learning the statistical analysis of measurement error. *International Journal of Science Education,* 15 (4), 427-438.

Song, J. and Black, P. (1992) The effect of concept requirements and task contexts on pupils' performance in control variables. *International Journal of Science Education*, 14 (1), 83-93.

Strauss, A. and Corbin, J. (1990) *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park: Sage.

Taylor, B. and Kuyatt, C. (1994) *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. National Institute of Standards and Technology: Technical Note 1297. Also available at http://physics.nist.gov/Pubs/guidelines/contents.html

Tiberghien, A., Veillard, L., Le Marechal, J-F., Buty, C. and Millar, R. (2001) An analysis of labwork tasks used in science teaching at upper secondary school and university levels in several European countries. *Science Education*, 85 (5), 483-508.

Toh, K. A. and Woolnough, B. E. (1990) Assessing, through reporting, the outcomes of scientific investigations. *Educational Research*, 32 (1), 59-65.

Tomlinson, J., Dyson, P. and Garratt, J. (2001) Student misconceptions of the language of error. *University Chemistry Education*, 5 (1), 1-8.

Van Heuvelen, A. (1991) Learning to think like a physicist: A review of research-based instructional strategies. *American Journal of Physics*, 59 (10), 891-897.

White, R. T. (1996) The link between the laboratory and learning. *International Journal of Science Education*, 18 (7), 761-774.

# Appendix I

# The probes in full

This appendix contains the front cover sheet pasted on the brown envelopes and probe question sheets used in both questionnaires.

Any changes made to probe questions used in the post study are shown in square brackets. The line of text in the square brackets replaces the corresponding line of text as used in the pre-study.

Comments in curly brackets indicate probes which are only included in, or omitted from, the post questionnaire.

| SURNAME: | FIRST NAME: | |
|---|---|---|
| D/ [E] | University of Cape Town<br>Department of Physics | Unique questionnaire number stamped here |

# Laboratory Procedures Questionnaire

## Instructions:

Write your name in the box above.
Inside this envelope there are pages numbered up to page 10.
Read the text below and answer the questions on each sheet.
If you need more space for your answers, then use the backs of the sheets.
It should take you between 5 and 10 minutes to answer each question.

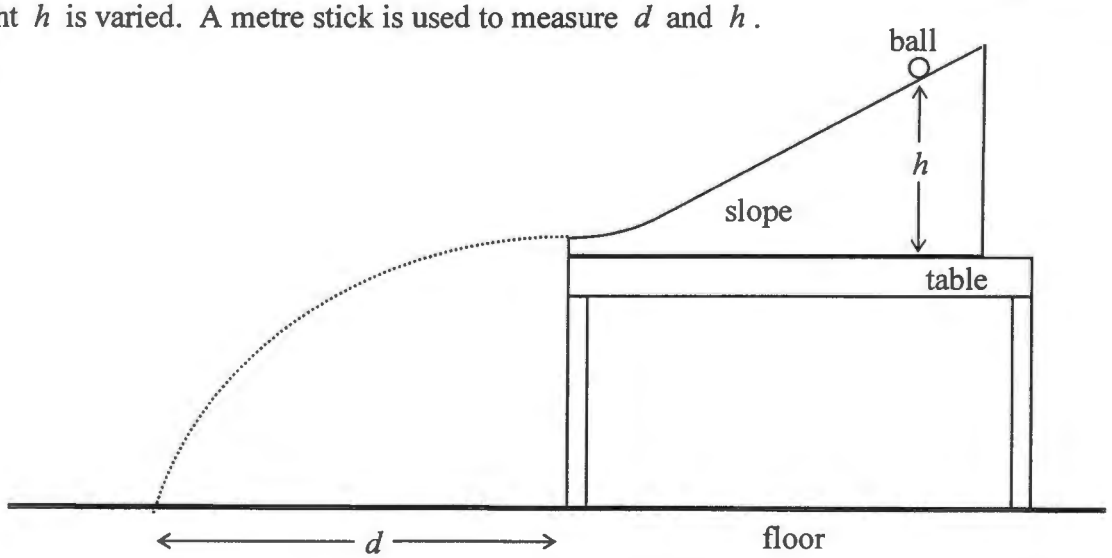
**Answer the questions in order and do not skip any sheet.**
**When you have completed a question, put the sheet inside this envelope and do not take it out again, even if you want to change your answer.**

**Note: It is possible that some answers may be similar or exactly the same as others. Please write all answers out in full, even if you feel that you are repeating yourself.**

## Context:

An experiment is being performed by students in the Physics Laboratory.
A wooden slope is clamped near the edge of a table. A ball is released from a height $h$ above the table as shown in the diagram. The ball leaves the slope horizontally and lands on the floor a distance $d$ from the edge of the table. Special paper is placed on the floor on which the ball makes a small mark when it lands.

The students have been asked to investigate how the distance $d$ on the floor changes when the height $h$ is varied. A metre stick is used to measure $d$ and $h$.



Unique questionnaire number stamped here

Q 1. (RD/D) [ Q1. (RD/E) in post questionnaire ]

The students work in groups on the experiment. Their first task is to determine $d$ when $h = 400$ mm. One group releases the ball down the slope at a height $h = 400$ mm and, using a metre stick, they measure $d$ to be 436 mm.

The following discussion then takes place between the students.

I think we should roll the ball a few more times from the same height and measure $d$ each time.

Why? We've got the result already. We do not need to do any more rolling.

I think we should roll the ball down the slope just one more time from the same height.

A B C

With whom do you most closely agree? (Circle ONE):

| A | B | C |
|---|---|---|

Explain your choice.

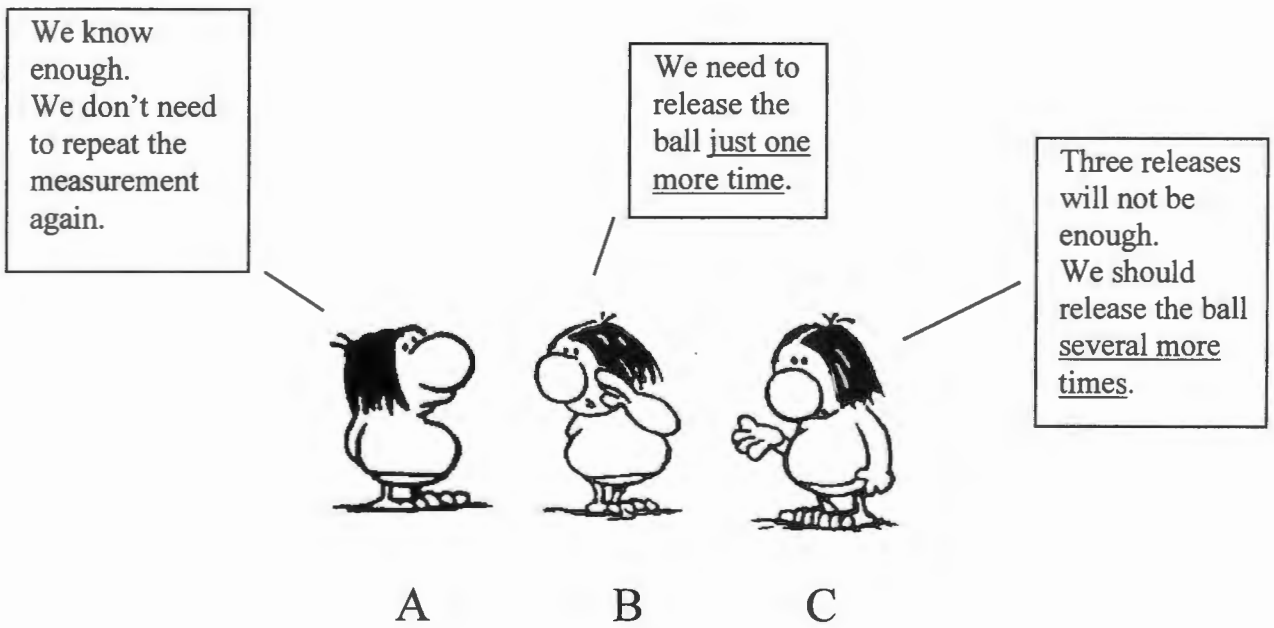
Unique questionnaire number stamped here

Q 2. (RDA/D) [ Q2. (RDA/E) in post questionnaire ]

The group of students decide to release the ball again from $h = 400$ mm.
This time they measure $d = 426$ mm.

| | | |
|---|---|---|
| First release: | $h = 400$ mm | $d = 436$ mm |
| Second release: | $h = 400$ mm | $d = 426$ mm |

The following discussion then takes place between the students.




With whom do you most closely agree? (Circle ONE):

| A | B | C |
|---|---|---|

Explain your choice.

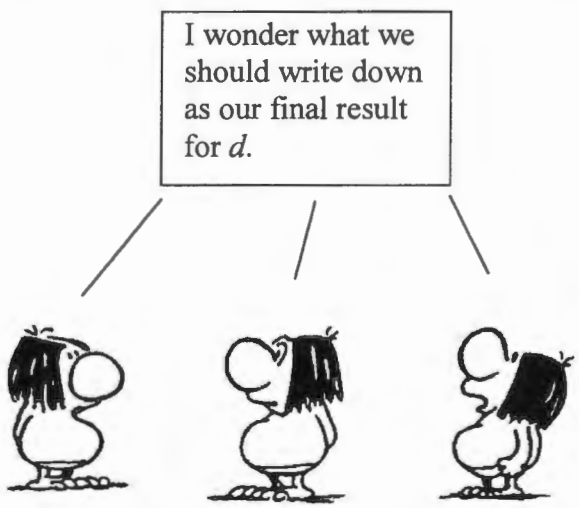
Unique questionnaire number stamped here

Q 3. (UR/D) [ Q 4. (UR/E) in post questionnaire ]

The students continue to release the ball down the slope at a height $h = 400$ mm.
Their results after five releases are:

| Release | $d$ (mm) | |
|---|---|---|
| 1 | 436 | |
| 2 | 425 | [426 in post questionnaire] |
| 3 | 440 | [438] |
| 4 | 425 | [426] |
| 5 | 434 | |

The students then discuss what to write down for $d$ as their final result.




Write down what you think the students should record as their final result for $d$.

Explain your choice.

Unique questionnaire number stamped here

Q 4. (AN/D) [ Q 5. (AN/E) in post questionnaire ]

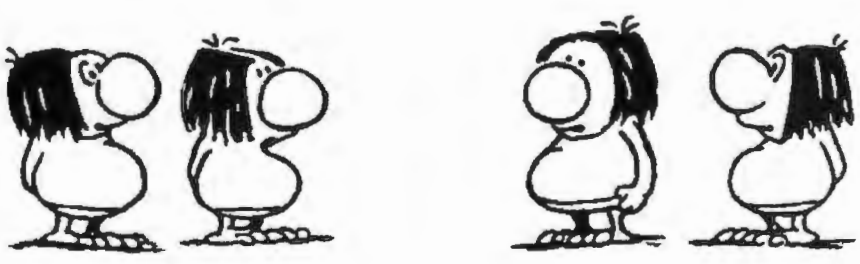
Another group of students have decided to calculate the average of all their measurements of $d$ for $h = 400$ mm. Their results after six releases are:

| Release | $d$ (mm) |
|---|---|
| 1 | 443 |
| 2 | 422 |
| 3 | 436 |
| 4 | 588 |
| 5 | 437 |
| 6 | 429 |

The students then discuss what to write down for the average of $d$.

All we need to do is to add all our measurements and then divide by 6.

No. We should ignore $d = 588$ mm and then add the rest and divide by 5.



A B

With whom do you most closely agree? (Circle ONE):

| A | B |
|---|---|

Explain your choice.

Unique questionnaire number stamped here

Q 5. (SMDS/D) [ Q 6. (SMDS/E) in post questionnaire ]

Two groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm. Their results for five releases are shown below.

| Release | Group A $d$ (mm) | Group B $d$ (mm) |
|---|---|---|
| 1 | 444 | 441 |
| 2 | 432 | 460 |
| 3 | 424 | 410 |
| 4 | 440 | 424 |
| 5 | 435 | 440 |
| Average: | **435** | **435** |

Our results are better. They are all between 424 mm and 444 mm. Yours are spread between 410 mm and 460 mm.

Our results are just as good as yours. Our average is the same as yours. We both got 435 mm for $d$.

I think the results of group B are better than the results of group A.

A B C

With which group do you most closely agree? (Circle ONE): | A | B | C |

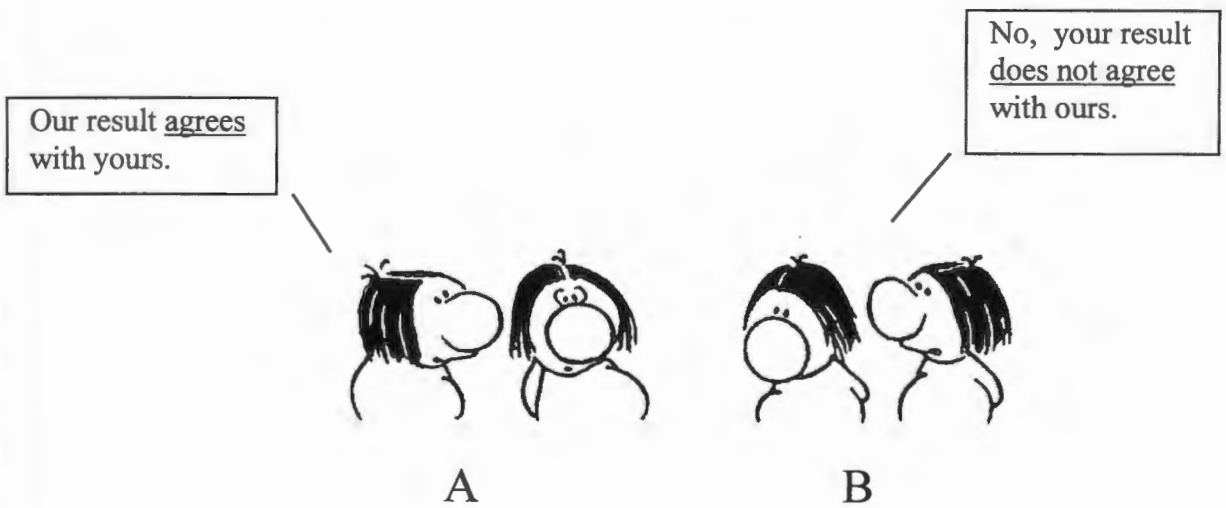
Explain your choice.

Unique questionnaire number stamped here

Q 6. (DMSS/D) [ Q 7. (DMSS/E) in post questionnaire]

Two other groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm. Their results for five releases are shown below.

| Release | Group A $d$ (mm) | Group B $d$ (mm) |
|---|---|---|
| 1 | 440 | 432 |
| 2 | 438 | 444 |
| 3 | 433 | 426 |
| 4 | 422 | 433 |
| 5 | 432 | 440 |
| Average: | **433** | **435** |




With which group do you most closely agree? (Circle ONE): A B

Explain your choice.

Unique questionnaire number stamped here

Q 7. (RT/D) { not used in post questionnaire }

The students are now given a stopwatch and are asked to measure the time that the ball takes from the edge of the table to hitting the ground after being released at $h = 400$ mm.
They discuss what to do.

We can roll the ball once from $h = 400$ mm and measure the time. Once is enough.

Let's roll the ball twice from height $h = 400$ mm, and measure the time for each case.

I think we should release the ball more than twice from $h = 400$ mm and measure the time in each case.

A B C

With whom do you most closely agree? (Circle ONE):

| A | B | C |
|---|---|---|

Explain your choice.

Unique questionnaire number stamped here

Q 8. (DMOS/E) { only administered in *post* questionnaire }

Two groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm. Their results for five releases are shown below.

| Release | Group A $d$ (mm) | Group B $d$ (mm) |
|---|---|---|
| 1 | 444 | 458 |
| 2 | 435 | 438 |
| 3 | 424 | 462 |
| 4 | 440 | 449 |
| 5 | 432 | 443 |
| Average: | **435** | **450** |

Our results agree with yours.

No, your results do not agree with ours.

A B

With which group do you most closely agree? (Circle ONE): A B

Explain your choice. Do not use the word "results" in your explanation.

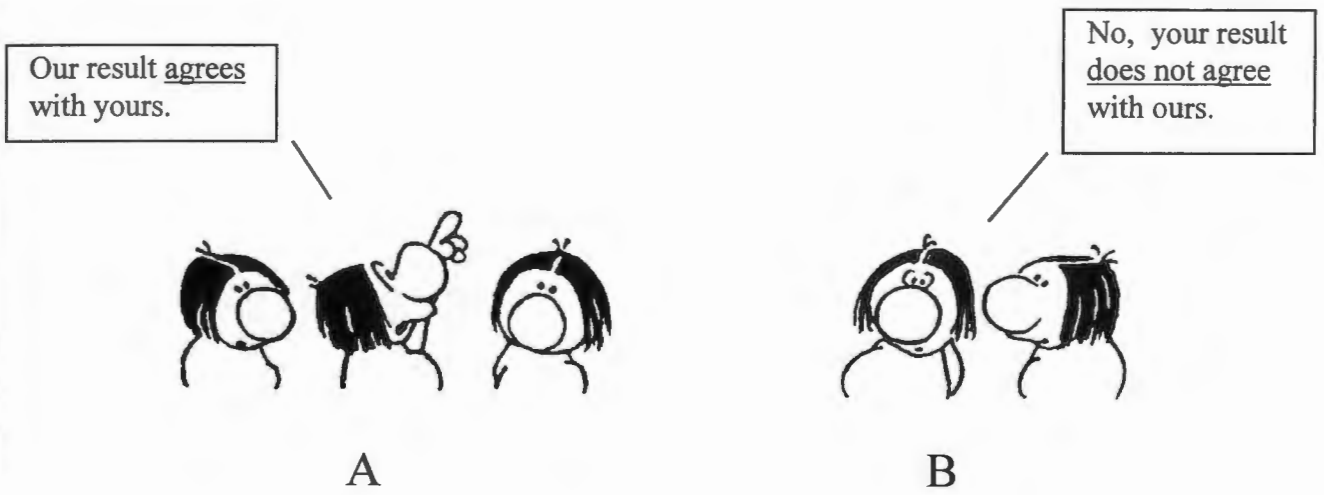
Unique questionnaire number stamped here

Q 9. (DMSU/E) { only administered in *post* questionnaire }

Two other groups of students compare their results for $d$ obtained by releasing the ball at $h = 400$ mm. Their means and standard deviation of the means for their releases are shown below.

Group A: $d = 436 \pm 5$ mm

Group B: $d = 442 \pm 5$ mm




A B

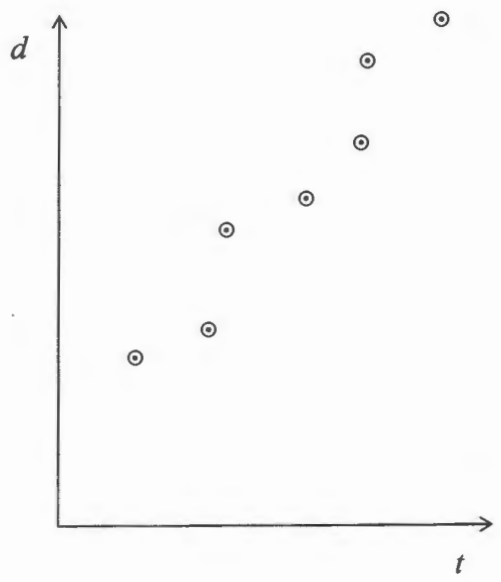
With which group do you most closely agree? (Circle ONE): | A | B |

Explain your choice. Do not use the word "result" in your explanation.

Unique questionnaire number stamped here

Q 8. (SLG/D) [ Q 14. (SLG/E) in post questionnaire ]

A group of students collect data at different heights and use it to plot a straight line graph. The data are plotted below. On this graph, draw the line that you think best fits this data.




Explain carefully what you have done and why.

Unique questionnaire number stamped here

Q 9. [ Q 15. in post questionnaire ]

Comments.

Are there any answers to the previous question sheets that you want to change?
**Please do not remove any sheets from the envelope.**
What was the question about and how do you want to change your answer?

Any other comments?



| In this laboratory questionnaire, I thought that the cartoon figures were (tick one): | male | female | mixed gender |
|---|---|---|---|

Unique questionnaire number stamped here

Finally, please fill in these details: {this page not included in post questionnaire]

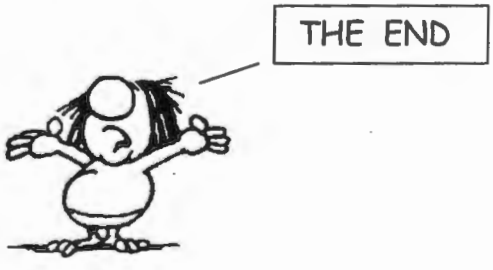
| **Surname:** | **First names:** |
|---|---|
| **Age:** | Circle one: **Male** **Female** |
| **Home language:** | **Second language:** |
| **Matric province:** | **Name of School:** |

Tick the **subjects** that you did in **matric**. Enter HG or SG and your symbol.
If you did subjects that are not listed, write them in the spaces provided.

| **Subject** | **Tick** | **HG / SG** | **Symbol** |
|---|---|---|---|
| English first language | | | |
| English second language | | | |
| Mathematics | | | |
| Physical Science | | | |
| Biology | | | |
| History | | | |
| Geography | | | |
| Afrikaans | | | |
| | | | |
| | | | |

Which programme have you registered on:

**Student number:**
(If you know it)



# Appendix II

# Tables of probe codes

*Table II.1: The complete set of codes assigned to individual students' responses to all the probes of the pre-instruction Laboratory Procedures Questionnaire (version D).*

| Set No. | Student No. | RD | RDA | RT | UR | SLG | SMDS | DMSS |
|---|---|---|---|---|---|---|---|---|
| 1004 | Stu001 | UA41a | UC45d | UC41ea | SF70 | SCF | SA42a | PA45 |
| 1078 | Stu002 | SA25 | UC65 | SC20 | SA60 | SLFO | PB21 | PA21 |
| 1062 | Stu003 | PC30 | SC81d | SC22sd | SA20 | SLFX | SA45 | PA43 |
| 1066 | Stu004 | SA25 | SC25 | SC25 | SA60 | PCM | PB23 | PA20 |
| 1051 | Stu005 | SA22 | SC24 | SC25s | SA64 | PLBO | SA25 | PA24x |
| 1071 | Stu006 | SA22a | SB25e | SC22 | SA22 | PLMO | SA43a | PA40 |
| 1048 | Stu007 | PC38x | SC25x | SC22s | SA71 | SLFO | PB65 | PA20 |
| 1017 | Stu008 | SA21ea | SC25d | SC28a | SA20 | PLBO | SA24 | PB30 |
| 1018 | Stu009 | PC30 | SC20 | UC60 | SA80 | SLF | SA10 | PB20 |
| 1037 | Stu010 | SA22a | SB25t | UC65e | SA22a | SLFO | SA43a | PA20 |
| 1067 | Stu011 | UA62 | UC60 | SC20 | SA20 | SLF | PB20 | PB61 |
| 1059 | Stu012 | SA22a | UC41a | UC45s | SA21 | SLFO | SA25 | PA22 |
| 1012 | Stu013 | UA65 | SC20d | SC25 | SA20 | SLFO | PB23p | PA22 |
| 1079 | Stu014 | SC82x | SB82ed | UC00 | SA70 | SLF | SA41a | PA20 |
| 1027 | Stu015 | SA22a | SC25d | SC25s | SA28 | SLF | SA41 | PA20 |
| 1040 | Stu016 | SA21 | SC20 | SC20d | SA25 | PLMO | PB23 | PA20 |
| 1074 | Stu017 | PC35p | PB35d | SC25 | SA22 | SLF | SA45 | PB25 |
| 1061 | Stu018 | SA82a | SC70d | SC70e | SA62 | SLF | SA42 | PA22 |
| 1060 | Stu019 | PC30t | PB31 | UC65 | SA25b | SCFO | PB10 | PA20 |
| 1047 | Stu020 | SA22a | UC42d | UC45 | SC22 | SLF | SA74 | PA30 |
| 1033 | Stu021 | SA21 | SC25d | UC65 | SA80 | PLM | PB40 | PA20 |
| 1024 | Stu022 | SA21 | UB63t | SC21e | SA60 | SLF | PB23a | PA20 |
| 1068 | Stu023 | SA22a | UC40d | UC41as | SA20 | SLFO | SA43a | PA22a |
| 1083 | Stu024 | PB01 | PC30d | PA30 | PE34 | PCX | PB20 | PB20 |
| 1050 | Stu025 | PC32 | PB30 | SC20s | PD81 | PLBO | PB21 | PA22 |
| 1082 | Stu026 | SA25 | SC25 | SC25 | SA25 | PLM | PB20 | PB20 |
| 1087 | Stu027 | SA21a | SC21a | PB30 | SA21ab | PLBO | PC41 | PA21 |
| 1049 | Stu028 | PB31 | SC22ad | PA35 | SA64 | SLF | SA14 | PA22 |
| 1009 | Stu029 | PB31 | PA61 | UC45 | UU00 | PPB | SA62 | PA42 |
| 1093 | Stu030 | SA25 | SC25 | SC25 | SA30 | PLMO | PB30 | PA33 |
| 1072 | Stu031 | SA25 | SC25 | SC22a | SA71 | PLMO | SA43a | PA30 |
| 1034 | Stu032 | PB31 | UC01 | UC60d | SA61 | SLF | SA40 | PA20 |
| 1046 | Stu033 | SA21a | SC25d | SC25s | SA25b | SLF | SA45 | PA28 |
| 1101 | Stu034 | SA25 | UC65d | SC25 | SA20 | SLF | SA75 | PA43a |
| 1102 | Stu035 | SA22p | UC63d | UC62s | SA32 | SLF | SA22 | PA43e |
| 1084 | Stu036 | SA21 | SB20d | SC22 | PD30 | SLF | SA40 | PA20 |
| 1023 | Stu037 | SA22! | UB80d | SC22 | SA20 | SLF | SA43 | PA22 |
| 1039 | Stu038 | SA25 | SB25 | SC25 | SA20 | SLF | SA20 | PB40 |
| 1021 | Stu039 | SA21p | UC65d | UC00 | SA31a | SLF | SA42a | PA25 |
| 1041 | Stu040 | SA21a | SC21 | SC21 | SA25e | SLF | PB21 | PA22 |
| 1057 | Stu041 | SA25 | SB25 | SC25 | SA25 | PLBO | SA45 | PA30 |
| 1100 | Stu042 | SA21 | UC40 | UC41ds | SA85 | SLF | PA35 | PA65 |
| 1086 | Stu043 | UA41a | SC81da | SC25 | SA25 | PLMO | SA25 | PB10 |
| 1058 | Stu044 | SC81e | PA45d | UC45s | SB22 | SLF | SA40 | PA25 |
| 1003 | Stu045 | SA21a | SC25d | SC21ea | SA70b | SLFO | SA15 | PA26 |
| 1045 | Stu046 | PC35 | SC81da | UC45d | SA80 | SCF | SA43aex | PB25 |
| 1001 | Stu047 | SA25 | SC25d | SC25 | SA20 | SLF | SA43a | PB20 |
| 1019 | Stu048 | UA42 | PB32 | UC42 | SA26 | SLFO | SA45 | PA22 |
| 1076 | Stu049 | SA22a | SC25 | SC22a | SA22a | SLF | SA45 | PA20 |
| 1089 | Stu050 | PB31! | SC20d | UC40d | SA70 | SLFO | PB10 | PA63 |
| 1010 | Stu051 | SA21a | SC25 | UC65 | SA25 | SLF | SA20 | PA20 |
| 1069 | Stu052 | SA21a | SC25d | UC40d | SA20 | SLFO | PB20 | PA80 |
| 1006 | Stu053 | SA21e! | UC65 | UC41s | SA81 | SLF | SA43a | PB80 |

*Table II.2: The complete set of codes assigned to individual students' responses to all the probes of the post-instruction Laboratory Procedures Questionnaire (version E).*

| Set no. | Student | RD | RDA | UR | SLG | SMDS | DMSS | DMOS | DMSU |
|---|---|---|---|---|---|---|---|---|---|
| 1441 | Stu001 | SA22ap | UC62p | SA20 | SLF | SA15p | PA40 | PB23xa | SA70 |
| 1432 | Stu002 | SA24a | SC24a | SA26 | SLF | PB41 | PA20 | PB20 | SA70 |
| 1412 | Stu003 | SA21e | SC24e | SA23a | SLF | SA70 | PA71σ | PB71σ | PB73 |
| 1444 | Stu004 | SA21a | SC25d | SA24 | SLF | SA70 | PA20 | PB20 | SA74 |
| 1427 | Stu005 | SA23a | SC23a | SA23b | SLF | SA40p | PB26 | PB26 | SA40 |
| 1458 | Stu006 | SA70ax | SC70d | SA23b | SLFO | SA43ap | PA22 | PB21 | PA23 |
| 1447 | Stu007 | UA41a | SC25d | SB21a | SLFO | SA43x | PA30 | PB40 | SA70 |
| 1403 | Stu008 | SA23a | SC23a | SA23 | SLF | SA70p | PA22 | PB71 | PB73 |
| 1404 | Stu009 | SA23a | SC72d | SA81 | SLF | SA70 | PB23 | PB27 | PB40 |
| 1440 | Stu010 | SA22 | SC22t | SA22 | SLF | SA70a | PB20 | PB20 | SA70 |
| 1455 | Stu011 | SA25 | SC25 | SA25 | SLF | SA40 | PB40 | PB20 | PB73 |
| 1454 | Stu012 | SA23 | SC23a | SA25 | SLF | SA40 | PA21 | PB20 | PA20 |
| 1428 | Stu013 | SA22 | SC23e | SA26 | SLF | SA20 | PB30 | PB10 | PB23 |
| 1459 | Stu014 | SA23a | SC70 | SA63 | SLF | SA73pm | PA30x | PB71σ | SA70 |
| 1406 | Stu015 | UA42 | UC40 | SA20 | SLF | SA40 | PA22x | PB22 | SA70 |
| 1463 | Stu016 | SA23x | UC41 | SA20 | PLM | PB20 | PA30 | PB22 | PA14 |
| 1435 | Stu017 | UA43 | UC64d | SA24 | SLFO | SA70m | PB26σ | PB22 | PB27 |
| 1420 | Stu018 | SA22 | UC42d | SA20 | SLF | SA42 | PA71σ | PB71σ | SA70 |
| 1413 | Stu019 | UA65 | UC45 | SA20 | SLF | PB25 | PA20 | PB20 | SA70 |
| 1445 | Stu020 | SA22b | UC65 | SA23 | SLF | SA73a | UB73 | UB73 | SA70 |
| 1462 | Stu021 | SA23b | SC23bd | SA23 | SLF | UB70 | SA70m | SB70 | SA70 |
| 1450 | Stu022 | SA24 | UB80te | SA23 | SLF | SA70e | PA21 | SB75σ | SA70 |
| 1410 | Stu023 | UA65e | UC42d | SA27a | SLF | SA43p | PA22 | PB20 | PB73 |
| 1426 | Stu024 | SA21 | SC22d | SA80b | PLM | PB20 | PB20 | PB20 | PB23 |
| 1431 | Stu025 | SA25 | SB20t | SA20 | PLMO | PB20 | PB22 | PB20e | PB23e |
| 1424 | Stu026 | SA20 | SB20 | SA20 | PLM | PB20 | PB20 | PB20 | PB27 |
| 1407 | Stu027 | SA23x | UC40dt | SA23 | SLF | SA10 | SA70x | SA70x | PA27 |
| 1423 | Stu028 | SA73a | UC60d | SA23 | SLF | SA40 | PA30x | PB21 | SA70 |
| 1448 | Stu029 | UA45e | UC45p | SA31 | SLF | SA42p | SA70e | SB70 | SB70 |
| 1466 | Stu030 | SA25 | SB70 | SA80 | PLM | SA40 | PA40 | PA40 | PA13 |
| 1461 | Stu031 | SA22a | SC22a | SA81a | SLFO | SA43a | PA20 | PB22 | PA23 |
| 1402 | Stu032 | SA24 | UC01 | SA61 | SLF | SA45 | PA27σ | PA20 | SA70 |
| 1405 | Stu033 | SA25 | SB25 | SA81 | SLF | SA20 | PA10 | PB40 | SA70 |
| 1417 | Stu034 | SA22a | UC65 | SA81 | SLF | SA70ep | SA70 | SB70 | SA70 |
| 1401 | Stu035 | SA22 | SC24ep | SA63a | SLF | SA40p | SA70 | SB70 | SA74 |
| 1453 | Stu036 | SA22a | SB25d | SA24 | SLF | SA70pm | PA10 | PB21 | PB23 |
| 1467 | Stu037 | SA25e | SC22 | SA23 | SLF | SA70 | PA30 | PB10 | PB27 |
| 1460 | Stu038 | UA42 | UC45 | SA61 | SLF | SA20 | PB20 | PB20 | PB26 |
| 1418 | Stu039 | UA62 | UC62t | SA60 | SLF | SA75 | SA70 | SA70 | SA70 |
| 1465 | Stu040 | SA22 | SC22 | SA20 | SLF | SA70m | PA40 | PB43σ | SA70 |
| 1452 | Stu041 | UA65 | SC25 | SA24a | SLF | SA44 | UB73σ | PB41 | SA70 |
| 1419 | Stu042 | SA20 | SC25 | SA60 | SLF | SA75 | SA70σ | PB27 | SB70 |
| 1436 | Stu043 | SA21 | SC22d | SA23a | SLF | SA75 | PB40 | PB22 | SA72 |
| 1456 | Stu044 | SA23 | SC24d | SA23 | SLF | SA70m | PA71σ | SB70σ | SA70 |
| 1422 | Stu045 | UA43a | UB45d | SA22a | SLFO | SA70 | PA10e | PB71 | PB73 |
| 1446 | Stu046 | UA65 | UB65d | SA23 | SLF | SA45e | SA70 | PB71 | SA70 |
| 1414 | Stu047 | SA25 | SC25 | SA20 | SLF | SA73a | PB28 | PB20 | PB73 |
| 1416 | Stu048 | PA30 | PC30 | SA20 | SLF | SA40 | PA20 | PB20 | SA70 |
| 1439 | Stu049 | SA22 | UB40d | SA60 | SLF | SA20 | PA20 | PB20 | SA70 |
| 1409 | Stu050 | SA75 | UC65 | SA23a | SLF | PB25 | PA45 | PB61 | SA70 |
| 1425 | Stu051 | SA25e | SC25e | SA60a | SLFX | PB20 | PA22 | PB20 | SA70 |
| 1411 | Stu052 | SA21a | SC22 | SA62 | SLF | PB20 | PA20 | PB11 | PB73 |
| 1451 | Stu053 | SA21p | SC72 | SA83 | SLF | SA70e | PB90 | PB20 | SA70 |