An Exploratory Study of Pre-service Primary Teachers' Understanding of Uncertainty in Measurements in Singapore

MD SHAHRIN K-S-MOORTHY

An Exploratory Study of Pre-service Primary Teachers' Understanding of Uncertainty in Measurements in Singapore

By Md Shahrin s/o K S Moorthy

First Supervisor: Ros Roberts
Second Supervisor: Phil Johnson

A Thesis Submitted for the Degree of Doctorate of Education

School of Education
Durham University
2015

ABSTRACT

This study was set in the context of a reform agenda for Singapore's science curriculum to adopt inquiry in teaching and learning science (MOE, 2008). Teachers, including pre-service primary teachers (PSTs) who were subjects of this study, are expected to engage their students with scientific evidence including measurements taken during science investigations. The inherent nature of measurements is that they are always affected by errors that caused uncertainty. Understanding this, as well as other procedural ideas underpinning uncertainty would be important for understanding evidence before looking at data that are subjected to uncertainties in measurements. Such understandings would be important for the PSTs when they teach their future students how to obtain valid and reliable data, and to evaluate the methods of investigation or scientific conclusions based on evidence.

This study, therefore, was aimed at exploring such understandings using the Concepts of Evidence (Gott, Duggan, and Roberts, 2008) as a theoretical framework. The lack of a research instrument customised to such a need motivated this study to develop one.

The study was carried out in two phases. The first involved fifty-five PSTs and directed towards getting an accurate interpretation of procedural ideas underlying uncertainty by triangulating the evidence from questionnaire and interviews and iteratively refining the "probes" as the study progressed. The second phase focused on developing a questionnaire based on findings from the first and testing it on twenty PSTs.

The results revealed that most PSTs could recognise uncertainties in measurements and suggest the right actions to deal with them, but they generally had difficulties explaining their actions implying shallow understanding

of concepts underpinning uncertainty, and reliance on routine knowledge. This has strong implications for teacher preparatory programmes as well as the teaching of procedural understanding.

DECLARATIONS

This thesis represents my work. No material contained in the thesis has previously been submitted for a higher degree in a university.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to those who have contributed their time and effort throughout the completion of this study.

First and foremost, I would like to thank my first supervisor, Ms Ros Roberts, for her untiring support, critical insights and for taking genuine interest in my studies and welfare, for without her guidance and sound advice I would not have been able to complete this thesis.

I would like to thank my second supervisor, Dr Phil Johnson, who has provided me with lots of helpful advice and thoughtful comments during the second phase of this research. His help and guidance in the writing of this thesis are very much appreciated.

Special thanks must also be given to Professor Richard Gott who supervised me in the early stages of my doctoral journey.

I am also thankful to my parents for the doors they have opened for me to view the world and their prayers for my health and well-being. I am also really grateful to my family-in-law and my brother plus his family for their constant support and generosity.

Finally, I am greatly indebted to my very understanding wife, Haliza, for her love, endless patience, and strong encouragement, and to my three children, Huzaifah, Yumni and Hazmi for their care and for reminding me about the most precious things in life.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS
(in alphabetical order)

Terms		**Abbreviations**
Concepts of Evidence	-	CofEv
Full-scale Deflection	-	FSD
National Institute of Education (Singapore)	-	NIE
Pre-service primary teacher	-	PST
Phase 1	-	P1
Phase 2	-	P2
Phase 1 Interview 1	-	P1I1
Phase 1 Questionnaire 1	-	P1Q1
Phase 1 Interview 2	-	P1I2
Phase 2 Questionnaire 2	-	P2Q2
Research Aims	-	RA
Research Questions	-	RQ
Standard Deviation	-	SD
Standard Error	-	SE
Variables (Controlled)	-	CV
Variables (Dependent)	-	DV
Variables (Independent)	-	IV

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Chapter Overview

The chapter begins by first examining the background of the research problem so that we can better perceive the situation that motivated this research. This is followed by a description of the research subjects: the pre-service primary teachers (PSTs[1]), followed by the context of the research problem, and a description of the types of science investigations carried out in local primary schools. Before the research aims and goals are stated, a few key terms will be explained. The research design comes next to explain why the qualitative approach has been adopted for the study. Finally, the chapter ends with an overview of the thesis.

1.2 Background to the Research Problem

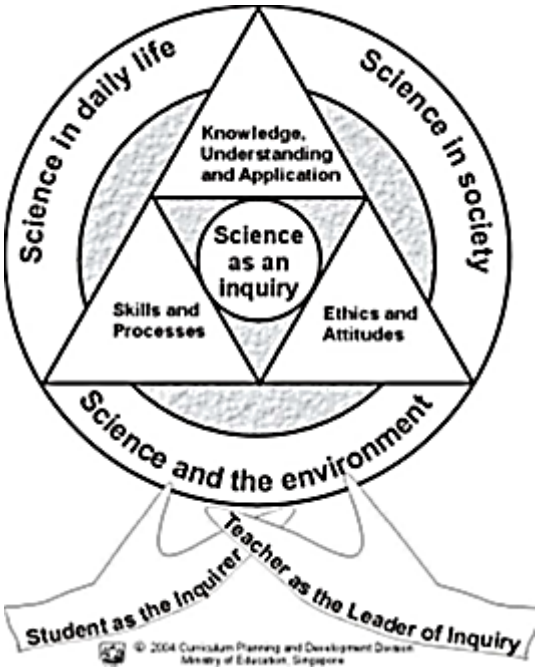
In 2001, Singapore's Ministry of Education (MOE, 2001, p.3) proposed that science should be taught by inquiry in order to develop skills and processes "to understand the natural world", and the acquisition of such skills and processes could be "realised primarily through the use of problem-solving exercises and practical investigations". However, in the few years following its introduction, there was little evidence to show that inquiry science had been implemented successfully (Chin & Kayalviszhi, 2002; Hogan et al., 2006; Poon, 2014).

In 2008, in response to economic imperatives, the education system revised its vision towards producing life-long learners and scientifically-literate citizens, and in tandem, the MOE reviewed its emphasis on inquiry science and

[1] To avoid confusion in this thesis, the term "students" will be reserved for all learners other than pre-service primary teachers; the latter will be referred by the abbreviation "PSTs".

the primary science syllabus was revised to become more inquiry-centric (MOE, 2008, see Figure 1.1).

Figure 1.1 MOE's Science Curriculum Framework (MOE, 2008, p.1)



The introduction of a new science curriculum framework with "science as an inquiry" at the core led to provisions of new curricular guidelines, training and resources in inquiry science for both in-service and pre-service teachers[2]. In order to provide greater clarity and alignment to its current goal, a new definition for inquiry that reflected the current emphases was also given:

> Scientific inquiry may be defined as the activities and processes which scientists and students engage in to study the natural and physical world around us. In its simplest form, scientific inquiry may be seen as consisting of two critical aspects: the *what (content) and the how (process) of understanding the world we live in.* (MOE, 2008, p. 11, words in italics are new emphases)

In translating its goals into action, MOE strongly recommended schools and teachers to use the five essential features of the inquiry classroom, which was originally proposed by the National Research Council (NRC) (2000), as the basic framework for an inquiry-based lesson (MOE, 2008). One feature exhorted teachers to provide students with opportunities to engage actively in

[2] The researcher was also involved in the development of teaching materials and training of in-service teachers during this period as part of his responsibilities at NIE (National Institute of Education, Singapore).

"the collection and use of evidence" (p.11), and MOE (2008) suggested that the main learning activity to achieve this should be *science investigations.*

The new definition for scientific inquiry calls for teachers to go beyond presenting the facts and outcomes of scientific investigations; students must also be shown the "how (process)" whereby the knowledge of scientific laws, concepts, and principles will be developed by actively engaging students with measurements. Teaching students to use measurements in this respect is more than how to take measurements; they must also be taught how to analyse and process data before interpreting the data as evidence. In addition, at every single step from taking measurements to interpreting data, the students must be taught how to evaluate data in terms of reliability and validity (Roberts & Gott, 2002; Warwick & Siraj-Blatchford, 2006; Duschl, Schweingruber, & Shouse, 2007).

MOE (2008) realised the goals of a new science curriculum cannot be achieved if teachers themselves do not embody the spirit of inquiry. Thus, returning to Figure 1.1, an important part of the framework is the vision for teachers to take on the role of a "leader of inquiry" This means, according to MOE (2008), the science teacher must go beyond their normal classroom instructional role and to "facilitate and role-model the inquiry process" (p.2). Words like "leader" and "role-model" suggest that a teacher in Singapore must be well-equipped with the "what (content)" and the "how (process)" as well in order to play an effective role in carrying out inquiry science in the classroom.

The pre-requisite of a science teacher in terms of knowledge and competencies has deep implications in teacher preparatory programmes. If there is a requirement to train PSTs to handle measurements then we need to understand their current state of knowledge and skills in that area. This is really

the crux of the research problem that motivated this study. But in order to better comprehend the research problem, some background knowledge about our research subjects - the PSTs, will be essential.

1.3 Pre-service primary teachers(PSTs)

Between 2006 and 2010, I had the privilege of being seconded to the NIE[3] as a Teaching Fellow, and one of my teaching duties was to prepare PSTs in the areas of pedagogy and curriculum development in primary science. The PSTs might be those attending the four-year NIE Bachelor of Science (Education) or Arts (Education) programmes, or the two-year Diploma-in-Education programme. After completing their programmes, the PSTs would be posted to primary schools to teach three subjects: English Language, Mathematics and Science. There was no subject specialisation.

Admission to the NIE programmes could either be based on a General Certificate of Education (GCE) Advanced-level Certificate or a Polytechnic Diploma. There was no preferred academic background for the GCE A-level certificate holders; the PSTs could have come from any academic stream: Arts, Commerce, or Science. Likewise, those from the polytechnics could either have attended a science or a non-science course. Based on this understanding, we could assume the PSTs' level of scientific knowledge and skills varied quite considerably.

For the PSTs who were participants of this study, their knowledge of inquiry science might come from their past learning experiences in schools. But this was expected to be quite limited given that inquiry science was poorly implemented across the local education system during the years when the PSTs were in school (Hogan et al., 2006; Kim, Tan, & Talaue, 2013). In the NIE

[3] The NIE, which is within Nanyang Technological University, is the only institution responsible for teacher training and qualification in Singapore.

teacher preparatory programmes, the major emphases would be on developing facilitation skills and knowledge of designing inquiry-based lesson plans. To achieve these, the PSTs would be asked to develop and carry out inquiry lessons during their NIE coursework and school practicum. To my knowledge, there was no learning module in NIE dedicated to teaching PSTs explicitly how to handle scientific evidence. Learning opportunities presented to PSTs in this area were unplanned and incidental in nature, and would likely to take place as they were designing or carrying out the scientific investigations themselves or with students during their school practicum. Since science investigations provide the context for the research problem of this thesis, it is essential that we know the types of investigations found in the local primary science curriculum.

1.4 Science Investigations in Singapore's primary schools

Investigations have been defined as activities that "involve formulating questions or hypotheses, devising fair methods and carrying out those methods to find out answers to the questions or to verify the hypotheses" (MOE, 2008, p.8). Accordingly, the process of investigation specified by the same syllabus involved several key steps:

- constructing a question or hypothesis;
- identifying variables, and specifying the variables to be controlled;
- devising a method to test the hypothesis;
- deciding on measuring devices and data to be collected, and then carrying out the measurements;
- drawing inferences from data and/or observations;
- evaluating whether the data and/or observations support or refute the hypothesis.

The description given in the preceding paragraph seemed to fit that of a variable-based investigation (Gott & Duggan, 1995), and this is indeed the main type of investigation carried out in the local primary science curriculum (MOE, 2008)[4]. Such an investigation normally explores the relationship between two variables, specifically the independent variable (IV), which may be categorically differentiated or quantitatively manipulated, and the dependent variable (DV), which may change as a result. Often, a number of repeated measurements are taken for the DV. Other factors (or variables) can also affect the DV causing difficulties in establishing the relationship between the IV and DV, so these factors need to be identified and controlled in order to ensure a fair test. Such factors are known as the controlled variables (CV).

Based on the classification provided by Gott and Duggan (1995), the variable-based investigations carried out locally are mostly Type 1 and Type 2 (see Table 1.1).

Table 1.1 Variable-based investigations in the local primary science curriculum

Type	Independent variable(IV)	Dependent variable(DV)	Examples
1	categoric	continuous	• Find out which liquid has the highest boiling point • Find out which material is the best heat insulator
2	continuous	continuous	• Find how the period of a pendulum varies with its length • Find how the time taken for sugar to dissolve in water varies with temperature

To illustrate further the investigations carried out in local primary schools, the reader can refer to Annex 1.1 (an example of a Type 1 investigation, “Slide Along”), and Annex 1.2 (an example of a Type 2 investigation, “Spring Along”) taken from the “Guide to Teaching and Learning of Primary Science” (MOE, 2009). In “Slide Along”, students will be taking measurements of force (continuous DV) using a forcemeter to pull a toy up a ramp laid with different types of surfaces (categoric IV). In “Spring Along”, students will use a ruler to

[4] Henceforth, the term “investigations” used in the thesis refers to “variable-based” investigations done in the classroom and laboratory; it excludes fieldwork.

measure different lengths of a spring (continuous DV) extended by different weights (continuous IV) that are hooked onto the spring.

In general, and as seen in “Slide Along” and “Spring Along”, the measurements that will be taken are for the three different variables:

(a) The controlled variable (CV) (for e.g., the height of the ramp in “Slide Along” and the original length of the spring in “Spring Along”). The measurement of the CV is often taken once only if it is carefully done. However, if the investigation demands a higher assessment of accuracy, it may be repeated several times just to check whether the CV is constant and not fluctuating (for e.g., the length of the “unstretched” spring might change slightly during the investigation);

(b) The independent variable (IV), which is being manipulated by the investigator in a Type 2 investigation (for e.g., in “Spring Along”, theoretically[5], this will be the different weights that are used). Like (a), the measurement for each IV interval is normally taken once only; and

(c) The dependent variable (DV) in both Type 1 and 2 investigations that will change in response to the IV (for e.g. the measurements of force in “Slide Along”, and the measurements of length of spring in “Spring Along” stretched by the weights).

Distilling from the preceding paragraphs (a) to (c), we can see that measurements in investigations are either taken singly or repeatedly; repeated measurements are generally taken to improve reliability as a result of the uncertainty in a single measurement. It is also important to reiterate the underlying objectives of performing scientific investigation are the development of the “how (process)” skills that will be needed by students to derive scientific

[5] In practice, the weight in grams can be found marked on the metal weights.

relationships and concepts (MOE, 2008). But what are these how (process) skills?

The Singapore's primary science syllabus (MOE, 2001) provided a list of a breakdown of these skills and processes, and what each could comprise (see Annex 1.3)[6]. Many of these skills and processes can be related directly or indirectly to measurements. Table 1.2 gives a summary of these skills and processes (extracted from Annex 1.3), and it gives a description of the possible objectives that PSTs may have to relate to if they were planning and carrying out measurement activities during investigations. Table 1.2 can perhaps also served as the knowledge base of procedural concepts linked to measurements that PSTs in this study would be expected to have.

Table 1.2 Skills and processes *directly* related to measurements in primary science investigations (extracted from MOE, 2001)

Skills/Processes	This involves
Observing	…gathering information about objects or events by using instruments to extend the range of accuracy of observations, and making quantitative observations that are relevant to a particular investigation.
Measuring and using apparatus	…using measuring instrument/apparatus to conduct investigations, and includes knowing their functions and limitations, selecting appropriate apparatus when measuring, handling apparatus correctly, and recognising the variability/reliability of measurement and the need to repeat measurement.
Communicating	…conveying and receiving information in various forms such as charts, tables, graphs.
Analysing	…identifying patterns and trends in data, the variables that will affect the outcome of an investigation, the relationships between variables, and those aspects which make an investigation unfair, and specifying variables to be controlled.
Generating	…making predictions from data, drawing inferences or conclusions from quantitative observations, and giving reasonable explanations based on evidence.
Evaluating	…deciding on the accuracy of data obtain in an investigation.

Borrowing from Shulman (1987), teachers including the PSTs in this study obviously needed to have more than the specified procedural knowledge than their students. Perhaps, this is even more critical in an education system that calls for teachers to be "leaders of inquiry". With reference to science

[6] A similar list was provided to all PSTs, including the participants of this research, during their Curriculum and Pedagogy in Primary Science module.

investigations, procedural understanding of ideas such as uncertainty, experimental errors, fairness, accuracy, and precision in measurements will be essential if they were to take on such a role (see Watson & Wood-Robinson, 2002; Gott & Duggan, 2003; Sharp et al., 2012).

1.5 The Problem: Early observations of PSTs' understanding of measurements

According to Gott and Duggan (2003), teachers dealing with measurements ought to know the ideas about evidence that underpinned both validity and reliability, and should place high priority on obtaining valid and reliable measurements. Bearing these in mind, it started me thinking whether the PSTs under my charge were competent enough to handle measurements in their inquiry-based activities. I wondered if they had sufficient understandings to be "leaders of inquiry". My daily interactions with the PSTs during theory and laboratory-based activities indicated the PSTs were at different levels of understanding; while most were able to carry out measurements, they seemed to have difficulty applying procedural ideas to different measurement situations. Below are some observations of PSTs' decisions when handling measurements; they generally show the PSTs having difficulties in applying or synthesising ideas about evidence that underpinned validity and reliability:

(a) In choosing a measuring instrument, some were not concerned with its resolution of scale, and for them, any instrument could be used *accurately* as long as the particular quantity being measured was predicted to fall within the limits of its scale;

(b) Experimental errors might be seen as mistakes or blunders, and once corrected, the true values[7] would be found;

[7] A "true value" is one obtained by a perfect measurement (JCGM, 2008). Thus, it is a hypothetical value.

(c) Due to experimental errors, one measurement might not be sufficient to yield an accurate value; thus, it would be essential to repeat the measurement so that a mean value could be obtained. However, some PSTs did not plan to repeat their measurements in their investigation;

(d) The number of readings to be taken was decided based on practical factors like time and convenience; some PSTs would stop at a fixed number of repeats for the DV, like three or five, regardless of the high degree of variation seen in the set of measurements;

(e) Finding a mean value became mechanical and the rationale was unclear; there were instances when several PSTs insisted on showing the calculation of a mean value from a set of the same readings just to fulfil a routine practice;

(f) The concept of variation could be absent for some PSTs who showed they were satisfied only if the same reading was obtained several times during data collection; small variations were not tolerated;

(g) An investigation would normally be carried out by repeating the same DV measurement repeatedly for a fixed number of times for a particular IV value before proceeding to the next point; the PSTs were not looking for the relationship between variables or checking to see whether the range for the IV was appropriate (the relationship between variables was sometimes established using only a small range of IV values);

(h) When measurements obtained for the DV were plotted, every point were expected to lie on a straight line; points that were *slightly* off the straight line might be unacceptable and re-taken (leading to a waste of time); the idea of line of best fit was absent.

Most of the problems I encountered with the PSTs, including those exemplified above, might be associated with their understanding of uncertainty in measurements. I arrived at this conclusion after consulting my supervisors, talking to fellow teacher-educators at the NIE, and conducting a review of the relevant research literature. I was not able to confirm my suspicions with any available survey instrument that suited my purpose. I therefore decided to focus my research on exploring the PSTs' understanding of uncertainty in measurements, and to use this knowledge to construct a questionnaire. On hindsight, I suspect many of these difficulties could have risen from a science curriculum that was too focused on building laboratory routines and outcomes, and too little focus on the "thinking behind the doing" (Roberts & Gott, 2006) and designing investigations (Kim, Tan, & Talaue, 2013).

Before the research goals are stated, we need to look at how a few key terms are defined in this thesis. This will provide us with a better understanding of how they are applied in the rest of the thesis, but the meaning of these terms will be developed further as the thesis progresses.

1.6 The meaning of key terms

1.6.1 "Measurement"

The term "measurement" is used in this thesis in several ways. A search in the dictionary (Random House, 2001) revealed a "measurement" can be "a measured dimension", "the act of measuring" or "the extent or size ascertained by measuring". In Section 1.5 earlier, the different measurements of IV, DV and CV are in fact "measured dimensions". In investigations, these might be "quantifiable observations" a student has to analyse for validity and reliability in order to use as evidence. "The extent or size ascertained by measuring" represents the value of a quantity (for e.g., length, mass, and time)

with a unit (for e.g., metres, kilograms, and seconds) and together they allow the student to interpret, make comparisons, and eventually draw inferences (Abruscato & DeRosa, 2010).

In the literature, words like “datum”, “reading”, “value”, and “measurement” are all used synonymously to represent a “measured dimension”. This thesis shall adopt this common practice.

1.6.2 “Uncertainty in measurement”

Measurements are never perfect; they would always have some degree of uncertainty or doubts about the measurements as a result of “systematic and random errors”. Thus, whenever the phrase “uncertainty in measurement” is used in this thesis, it means there is a “margin of doubt” (Bell, 1999, p.1) over a “measured dimension” because of the effects of scientific errors. The size of this “margin of doubt” reflects the quality of the measurement, and it eventually informs about its validity and reliability.

It is important to note however that this thesis deals with the qualitative understanding of uncertainty; it is not concerned with the quantification or the calculation of uncertainty[8], but some statistical ideas concerning the estimation of uncertainty in repeated measurements will be included.

It is critically important that the PSTs in this study were able to understand the causes of uncertainty, which might be elusive, and if necessary in the context of their work, attempt to minimise them. From my preliminary observations of the PSTs, understanding uncertainty may be difficult. Next will be how the term “understanding” is understood in this thesis.

1.6.3 “Understanding”

According to the Revised Bloom’s Taxonomy (Krathwohl, 2002) that is widely referred to in the education context, the term “understanding” is more

[8] For example “error analysis” or estimation of uncertainty based on Type “A” or “B” evaluations.

than just “knowing” or “recalling”, understanding is about being able to construct meaning of knowledge (including procedural knowledge), as well as to provide explanation and the implication of its uses.

The term “understanding” that will be used in the context of this research has a more specific connotation; when applied to science investigation, it means “procedural understanding” (Gott & Duggan, 1995). A simple definition of this term is the understanding of scientific evidence that underpins the decisions made during the procedures employed in science (Gott, Duggan, & Roberts, 2008).

In order to provide a description of procedural understanding, Gott and Duggan developed a taxonomy using skills and the “concepts of evidence” as shown in Table 1.3.

Table 1.3 Procedural taxonomy (Gott & Duggan, 1995, p.34)

• Knowledge and recall of skills
• Understanding of concepts of evidence
• Applications of concepts of evidence (in unfamiliar situations)
• Synthesis of skills and concepts of evidence (in problem solving)

At this juncture, suffice to say that the term “concepts of evidence” is used by Gott and Duggan as well as others, to refer to concepts that are involved in the design of the task (for e.g., fair test), measurement (for e.g., choice of instrument), data handling (for e.g., use of tables and graphs) and the evaluation of the investigative task by checking the validity and reliability of the ensuing evidence. Importantly, the concepts of evidence include mathematical concepts associated with data analysis and processing as well as procedural ideas related to uncertainty in measurements. We shall see more of “concepts of evidence” in another chapter of this thesis.

Going back to the taxonomy, how do we apply this? To illustrate, if the knowledge and recall of skill refer to the correct use of a thermometer, then

understanding the concepts of evidence (for e.g., the concepts of accuracy, calibration, etc.) may involve figuring whether the chosen thermometer meets the level of accuracy. The application of concepts means applying the same procedural ideas (for e.g., accuracy, calibration, etc.) to different types of investigations thereby showing the ability to recognise how these ideas affect the quality of data and the resultant claim. Finally, synthesis can refer to the ability of using the same concepts in evaluating whether a set of reported temperature measurements is valid and reliable in the context of the whole investigation.

Having introduced the key terms, we shall next turn to the research aims and questions.

1.7 Statement of Research Aims and Research Questions

There are two interrelated aims in this research. First, it intends to explore the PSTs' understanding of uncertainty in measurements carried out in science investigations. By referring to Section 1.4, these measurements can be:

- A single measurement (for a CV, or an IV value in a Type 2 investigation);
- Repeated measurements (for a DV in a Type 1 investigation where the DV corresponds to a categoric IV);
- Repeated measurements (for a DV in a Type 2 investigation where the DV corresponds to a continuous IV).

Second, in order to study the PSTs' understanding of uncertainty in these measurements, observations alone may be insufficient; a more reliable and efficient way to see differences in the PSTs' understanding of uncertainty in the measurements could be via an instrument, a questionnaire. To develop items

for such a questionnaire, however, we first need to have a good understanding of the PSTs' ideas related to uncertainty in measurements.

Based on the preceding paragraph, the following therefore will be the research aims of this study:

1. To explore and describe pre-service primary teachers' understanding of uncertainty in measurement; and
2. To develop a questionnaire that allows me to see the patterns and divergences in the PSTs' understanding of uncertainty in measurement.

To achieve Research Aim 1 and Research Aim 2, the following questions will have to be addressed in this study:

(a) Do the PSTs expect the inherent variability of repeated measurements? Do the PSTs believe in true values? How do they understand the terms "accuracy" and "precision"?

(b) How do the PSTs choose their best measuring instrument to take a single "isolated" measurement (of a CV or IV)?

(c) What is the PSTs' purpose of repeating measurements? What do they think are the causes of variation in repeated DV readings? How do they decide on the number of repeats for a set of DV measurements? What procedural ideas are used when they select repeated data? What data characteristics were referred to by PSTs when they choose between two sets of data with an overlapping range? How do they handle anomalous result?

(d) How do the PSTs plan to take DV measurements in an investigation? How do they process "messy[9]" tabulated DV data from an investigation?

[9] as a result of uncertainties

(e) In the process of finding answers to (a) to (d), and designing a questionnaire that allows the PSTs to have a clear interpretation of its intended purpose, what does it reveal about the PSTs' ability to articulate their understandings of uncertainty in measurements?

1.8 Research Design

To the best of my knowledge, a study on PSTs' understanding of uncertainty in measurements has not been carried out in the Singapore context and there is no document either from the MOE (Singapore) or the local academia that clearly spells out the competence of teachers in this area. It might have been done elsewhere, but a search in the literature did not also reveal one. Because of this, the research conducted in this study will be exploratory in nature (Creswell, 2008); one that "seeks to find out how people get along in the setting under question, what meaning they give to their actions" (Schutt, 2012, p.13). The study will use qualitative methods in gathering and analysing the research data as these allowed different ways of exploring the rich understandings the PSTs might have, and the possibilities of clarifying and refining the interpretations of the PSTs' procedural ideas related to uncertainty in measurements.

1.9 Overview of Thesis

This thesis has eight chapters. Chapter 1 focuses on the research context and motivation. It also highlights the goals of this study. The literature review in Chapter 2 is devoted to four areas: first, to develop an understanding of the science of uncertainty; second, to look into methodological principles that could be used to develop a clear picture of the PSTs' understanding of uncertainty; third, to examine studies on uncertainty in order to gain insights into methodology and key findings; and finally, to establish a theoretical framework

that would guide this research in exploring the PSTs' procedural understanding of uncertainty in measurements. Chapter 3 is on methodology and the discussions will centre on the research approach, which consisted of two phases of studies, the sample, the methods of data collection, and the analysis of the instruments used in the study.

Chapters 4, 5 and 6 report the findings of different instruments used in the first phase mainly to address Research Aim 1, but by reviewing the design and the efficacy of the probes in each instrument, the chapters will also address Research Aim 2. Chapter 7 reports on the second phase of the research which centres on the development of the proposed questionnaire, and thus focusing mainly on Research Aim 2. Nevertheless, the results of the finalised questionnaire will also add to the findings from the earlier phase. Finally, the study concludes in Chapter 8 by discussing the implications and limitations of the research as well as some possible future directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Overview

There are four main emphases in this chapter. The first is to describe the science of uncertainty in measurements. Next, to develop methodological principles to explore the PSTs' understanding of uncertainty. Following that will be a review of research that investigated understanding of uncertainty in measurements. The review specifically seeks to study the methods and findings for the purpose of applying the learning to answering the research questions. The final part of the chapter will see it developing a theoretical approach for conducting its investigations.

2.2 The Science of Uncertainty in Measurement

2.2.1 Uncertainty in Measurement: defining it further

From Section 1.6.2, uncertainty is known to be inherent in the measurement of all variables and described as a "margin of doubt" caused by a combination of experimental errors shown by Equation (1).

$$\text{Error} = \text{Systematic error} + \text{Random error} \quad (1)$$

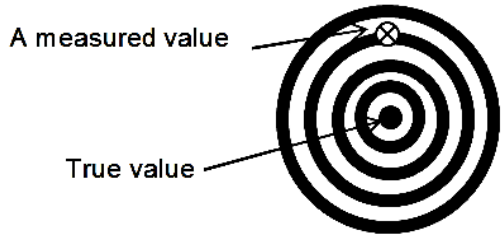
Additionally, uncertainty is created by the difficulty in establishing which of the errors dominate and the fact that the errors cannot be eliminated (Kirkup and Frankel, 2006). As uncertainty in measurement is often described in terms of "accuracy" and "precision", we shall look at these two terms next.

2.2.2 Accuracy and Precision

In principle, each measurement is taken to establish the true value of a measured quantity. If we are looking at a single measurement, an error is the gap between the measured and the true value (to illustrate, see Figure 2.1) and can be represented by Equation (2):

$$\text{Error} = \text{Measured value} - \text{True value} \quad (2)$$

Figure 2.1 Representation of an error for a single measurement



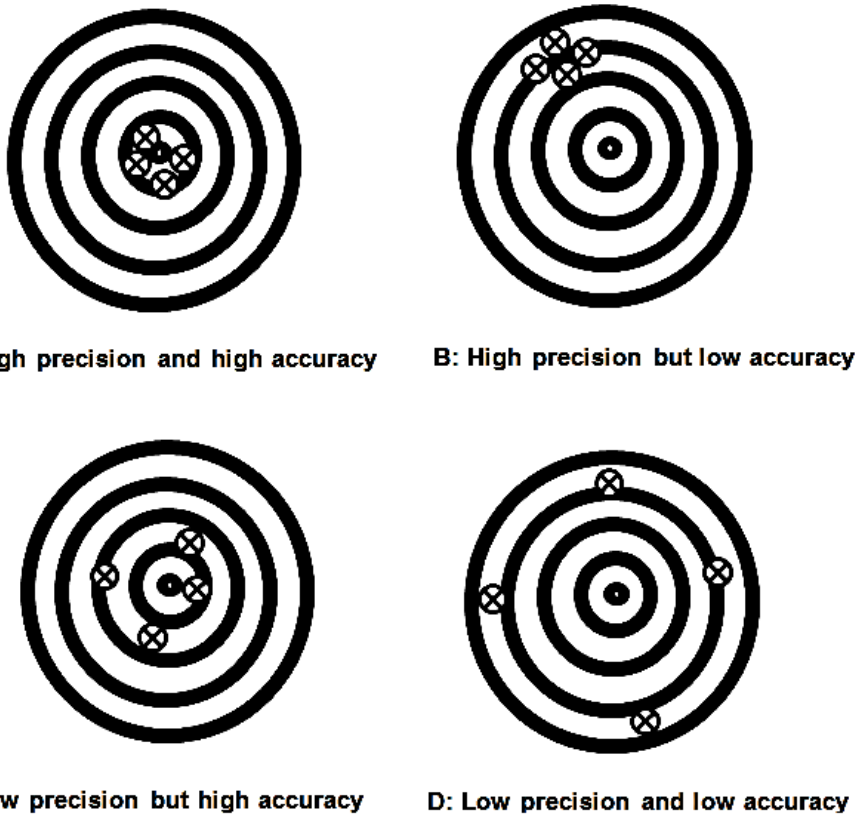
Based on Equation (2), accuracy can therefore be defined as the closeness of the measurement to the true value (Gott & Duggan, 2003). If a single reading is deemed close enough to the true value in the assessment of accuracy (and satisfies the purpose of the measurement), then repeating the measurement may not be necessary.

It is important to note, however, in the literature, many conceptual terms in measurements including accuracy have a multiplicity of meanings and this can lead to confusion. For instance, accuracy may be used for describing the correctness of choosing an instrument or method of measurement to indicate the value for a particular variable; if either of these is inappropriate (for e.g., choosing the girth of a tree to indicate its age will not be accurate), then the validity of the investigation can be called into question. Accuracy may also mean “trueness”, which is used for large sets of repeated readings and defined as a measure of the extent in which repeated readings of the same quantity give a mean that is the same as the true mean (Joint Committee for Guides in Metrology [JCGM], 2012). In this thesis, however, the term “trueness” shall not be used to avoid confusion over the use of too many equivalent terms.

What about precision? Precision is the degree of consistency and agreement among independent measurements of the same quantity (Gott & Duggan, 2003). Words like “spread”, “scatter”, “dispersion”, “variation” or “repeatability” have all been used to convey the idea of precision. When a set

of readings taken by a measuring instrument are similar or almost similar, we can say the data is precise, and the instrument that measures them reliable. If we used a similar diagram to the one seen in Figure 2.1 to illustrate precision and accuracy, we would end up with four possible situations shown in Figure 2.2.

Figure 2.2 The interrelationship between precision and accuracy for repeated measurements



A: High precision and high accuracy **B: High precision but low accuracy**

C: Low precision but high accuracy **D: Low precision and low accuracy**

Figure 2.2 shows a set of readings can be both precise and accurate (**A**), precise but not accurate (**B**), or the other way round (**C**), and finally, neither precise nor accurate (**D**). Basically, a set of readings would be accurate if the readings are clustered together close to the true value while precise readings show that all the readings are close to each other but not necessarily to the true value. As seen earlier, uncertainty in measurement is dependent on systematic and random errors, so we shall discuss these two terms next.

2.2.3 Systematic and Random Errors

A systematic error is one in which the true value is consistently overestimated (sometimes known as positive bias) or consistently underestimated (thus, negative bias). Systematic error is often very small and sometimes goes undetected, and this makes it potentially dangerous. Unlike random errors, a systematic error cannot be revealed by repeating a measurement under the same conditions with the same instrument, or its effects estimated by taking the mean value of several readings. At the primary level, systematic errors are normally identified by checking the instrument (for e.g., Vernier callipers) for “zero errors” or by calibrating the instrument (for e.g., a weighing balance) against a pre-defined standard.

Random errors affect the precision in a set of measurements; in science investigations, this may be manifested by the variation shown in the repeated data. The sources of random errors in investigations can come from uncontrolled variables, the characteristics of the measuring instruments themselves, and human errors in relation to both the control of variables and the use of instruments.

It is essential to note that systematic and random errors are defined according to whether they produced a systematic or random effect. We cannot say a certain source of error is inherently systematic or random; the same source of error may give rise to both effects. For instance, in operating a stopwatch in a pendulum investigation, we might not only start and stop in a slightly irregular manner when taking time, thus producing a random error, but we might also develop a tendency to always start too early or stop too late, which give rise to a systematic error. Although this being case, several sources of errors are well-established in primary science investigations and these will be

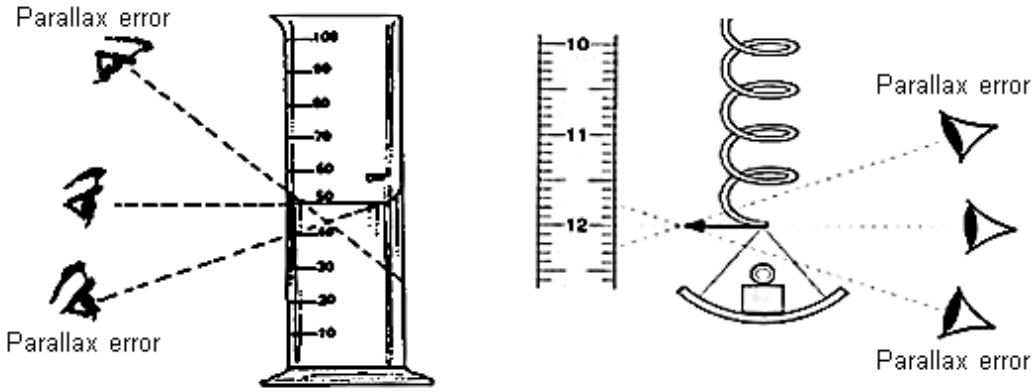
highlighted next to extend the idea of errors that contribute to uncertainty in measurements. The examples given are not intended to imply the level of understanding required of a PST in terms of the possible sources of errors, and neither do they attempt to represent all sources of errors in primary science investigations. Such an attempt will be futile as the sources of errors are just too numerous.

2.2.4 Common Systematic Errors in Primary Science Investigations

The sources of systematic errors can be traced back to human errors and the properties of a measuring instrument. In the following descriptions, we must bear in mind that the accuracy of a measurement is always affected:

- "Systematic human errors": one example is the reaction time in starting and stopping a stopwatch; the investigator may consistently delay starting the stopwatch after observing the start of an experiment or consistently stop the stopwatch before the end of the experiment. Another example is parallax errors (see Figure 2.3) whereby the investigator views the scale of a measuring instrument *consistently* at an angle rather than directly in front of it.

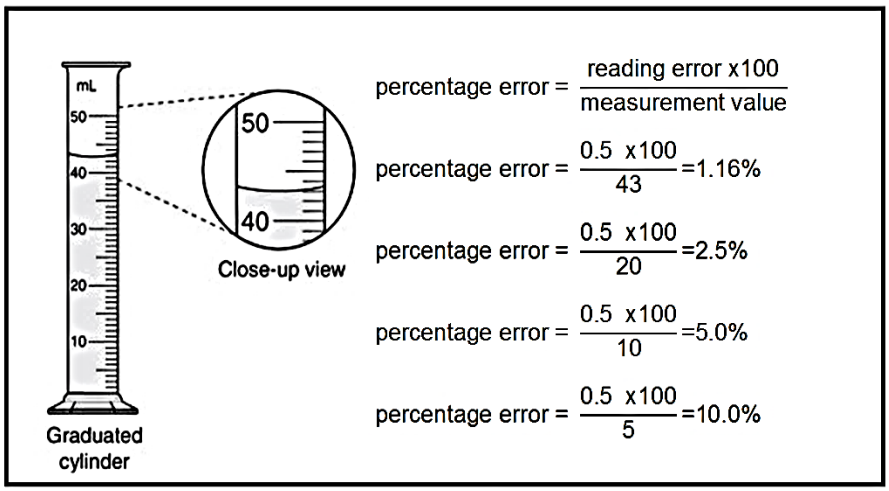
Figure 2.3 Parallax errors in reading a measuring cylinder and ruler (image modified from http://www.cnx.org and http://www.antonine-education.co.uk)



- "Zero errors": instruments (in primary science, for e.g. weighing balance, Vernier callipers, etc.) may develop a built-in error known as "zero error" over time, and therefore, must be checked to see if they are properly "zeroed" before the instruments are used. A voltmeter, for instance, may show a reading of 0.1V when it is not connected to an electrical circuit. So, when it is used to measure voltages, each measurement will have an error of 0.1V that should be deducted.
- "Reading errors": are due to the limitations associated with the resolution of scale, which is the smallest division on an instrument that can be read easily. Such an error becomes a concern whenever there is a need to interpolate a reading that happens to fall between two divisions. Thus, for a metre rule(smallest division: 1millimetre), it is reasonable to say that at best the length of an object can be read to the nearest millimetre, but to measure an object to greater accuracy, the rule of thumb allows us to use half the smallest division, thus for a metre rule, its reading error will be 0.5mm. Therefore, a reading that falls between 46 and 47mm can be written as 46.5±0.5mm.

  Sometimes, to measure a particular quantity, students may be confronted with having to choose the *best* instrument (in terms of giving the most accurate measurement) from a range of instruments with different scales. In such a situation, the best choice will probably be the instrument that measures the quantity nearest the end of its scale [the term "full-scale deflection (FSD)" is commonly used to represent the idea (Gott & Duggan, 2003)]. We can explain the reasoning using the reading error expressed as a percentage (see the example of a 50cm^3 measuring cylinder in Figure 2.4).

Figure 2.4 Percentage error of a 50cm^3 measuring cylinder (image taken from http://chemwiki.ucdavis.edu)



Let us suppose the best we can read the divisions on a 50cm^3 (smallest division: 1cm^3) measuring cylinder is to within 0.5cm^3; the best estimate therefore for any reading is measurement±0.5cm^3. Looking at the calculations in Figure 2.4, they indicate the reading error of 0.5cm^3 is relatively more significant at lower readings than at higher readings. To illustrate, ±0.5cm^3 represents only a 1.16% error for a 43cm^3 measurement, but it becomes a much larger error of 10.0% when a smaller quantity such as 5.0cm^3 is measured using the same measuring cylinder. It is this percentage error that really matters; the lower it is the better; thus, for a 5.0cm^3 quantity, it will be best measured by a 10.0cm^3 measuring cylinder (reading error=0.1cm^3; percentage error=2.0%) assuming it is the only other measuring cylinder available.

- "Readability": the term "readability" is often used in instruments with digital scales and it refers to the smallest change in mass that corresponds to a change in displayed value. For instance, an object of mass 154.348g when weighed on a scale with 0.01g readability only reads "154.35g". Since digital or electronic instruments can only display

up to a certain number of significant figures, this introduces uncertainty as the last digit will be a rounding-up or rounding-down number.

2.2.5 Common Random Errors in Primary Science Investigations

Random errors are due to human errors, uncontrolled variables, the non-reproducibility of method of measurement, and the characteristics of a measuring instrument (an instrument may rely on the conversion of the variable being measured into one that is easily read, for instance in a thermometer, the measured temperature is first converted to a change in volume, and then to a change in the length of the mercury thread. The conversions, however, may not be consistent every time it happens). We shall look at each one but bearing in mind they all affect the precision in a set of measurements.

- "Random human errors": these are errors that are committed inconsistently or unconsciously by investigators. Such errors cannot be completely removed by simply adopting the correct procedure; to a certain extent, they can be reduced by conducting a few preliminary trials that allow the investigator to practise the procedure, and to get a "feel" for the instruments. A good example of such an error is the inconsistent "reaction times" in starting and stopping a stopwatch for time measurements. Another is the inconsistency in deploying a measuring instrument. For instance, when using the thermometer, the bulb may be resting at the bottom of a beaker for one reading, but at the centre when the next reading is taken.
- "Uncontrolled variables": besides the IV, other variables can affect the measurements of a DV; slight changes that occur in such variables (sometimes, intermittently) during an investigation can result in variations in the measured data. These factors must therefore be

identified and controlled so that a fair test can be achieved. We normally assume the controlled variables(CV) will remain constant, but in reality, this may not happen. To illustrate, in the bouncing ball investigation, the ball was supposed to be released from a fixed height; instead, it might be released unconsciously from slightly different heights, higher or lower than the proposed one. This could result in tiny differences in the DV(rebounding heights).

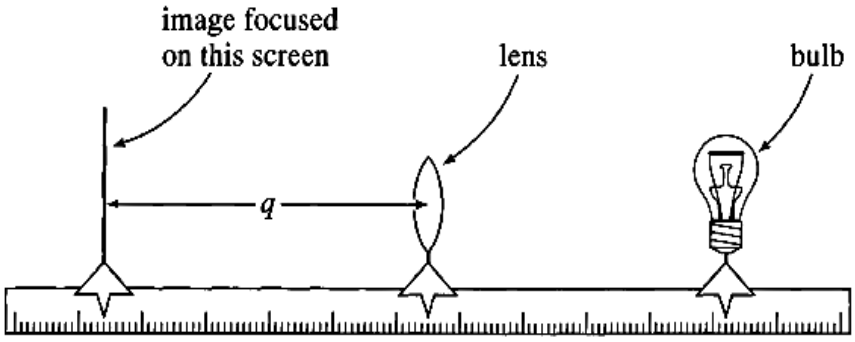
There is also a possibility that some of these factors were not or could not be identified, and therefore, left uncontrolled. For instance, the ball could have been thrown with some force unknowingly or it might have landed on a spot which has a different texture or hardness, or there might be fluctuations in the environment (draughts from outside the laboratory affecting the motion of the ball in air or changes in the room temperature that might alter the elastic nature of the ball). The possible sources of errors from uncontrolled variables are just limitless.

- "Instrument reliability": an instrument may have certain limitations (not faults) that do not allow it to measure consistently what it is supposed to. This can be due to its internal operating mechanism; for instance, a weighing balance may be affected by environmental conditions (for e.g., humidity, temperature) that prevent it from giving the same weight for a particular object. The technical term that is often associated with instrument reliability is "repeatability" and it refers to the ability of an instrument to bring about the same successive measurements of a particular quantity using the same method at the same location over appropriately short intervals of time (Gott, Duggan & Roberts, 2008).

- "Non-reproducibility of the methods of measurement": sometimes, it is not possible to make exactly the same measurements because the method is not properly defined. For instance, two investigators measuring the length of a rope may get different results because each may be stretching the rope with different amounts of tension or the rope could be frayed at its ends, thus making it difficult for the investigators to decide which ends they should be looking at.

  The problem of definition has also been illustrated by Taylor (1997) using Figure 2.5.

Figure 2.5 Uncertainty in an optic investigation (from Taylor, 1997, p.48)



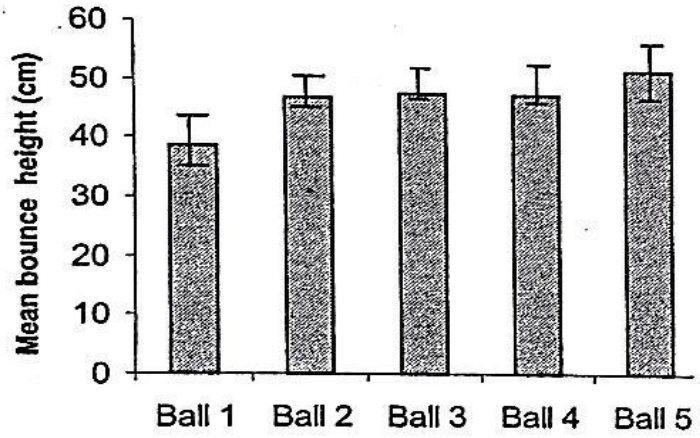
Taylor claimed it would be difficult to measure the actual distance between the lens and the image focused on the screen (length "**q**") as the lens was several millimetres thick and locating its centre was not easy especially when it was mounted on a bulky lens holder. Besides, the image formed on the screen might be well-focused over a range of many millimetres.

Another example concerns taking "dynamic" measurements like the rebound heights in the bouncing ball investigation. If the heights were taken against a metre rule using the naked eye only, the heights should only be noted at the instance when the ball reached the maximum height; but sometimes unconsciously, the heights were taken just

before the ball reached the top or on its way down, in such cases, a "reading error" would also be introduced.

The size of such an error can be a factor in deciding whether a particular method is "good enough" (Gott & Duggan, 2003, p.130). If the amount of uncertainty in measurements (say caused by reading errors) was large, the measurement procedure might not be good enough to address its intent. To illustrate, see Fgure 2.6; it shows a bar chart based on an investigation to find out whether a set of five balls (categoric IV) were similar in terms of their "bounciness" (expressed by the mean bounce heights).

Figure 2.6 Bar chart showing the mean bounce heights of five balls (taken from Gott & Duggan, 2003, p.131)



We cannot tell the differences between the five balls in terms of their rebound heights since their mean bounce heights fell within the overlapping error bars (which represented the variations in height measurements). The method of measurement used in the investigation was therefore not good enough because it produced large errors resulting in high degrees of variation that "concealed" the effects of the IV on the DV.

Having described random errors, the next section looks at statistical ideas to estimate such errors and explain the purpose of repeated measurements.

2.2.6 The Statistical Descriptions of Repeated Measurements

The simplest way of reporting variation in a data set is to use the concept of range. As an indicator of dispersion, it is based on the distance between the highest and lowest reading but this may be misleading as it is affected by abnormal data that can be exceptionally high or low. To avoid such false impression, we can use inter-quartile range, which quotes a fifty per cent central range that covers half of all measured values thereby making it more representative of the majority. But then again, like range, it is also based on two values (the 25th and 75th percentiles) of the whole data set.

Statistically, a more powerful measure of dispersion that takes into account *every value* in a data set will be standard deviation (SD) shown by Equation (3) below:

$$\mathrm{SD} = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \qquad (3)$$

Where SD = standard deviation
$\sum$ = sum of
x = a single measurement
$\bar{x}$ = mean value
n = number of measurements/sample size

An important condition to use SD is the sample of measurements must have a "normal distribution". What is normal distribution? If we plot a frequency distribution for *n* number of measurements[10], which is a random sample taken from a whole population of similar measurements (*N* being the total number), then we may get a histogram as shown in Figure 2.7a. If we keep increasing the size of the sample, the histogram will gradually take the form of a bell-shape curve (Figure 2.7b). A very large number of measurements allows us to make a fine subdivision of the scale and the histogram becomes a continuous curve known as the "normal (or Gaussian) distribution" (Figure 2.8).

[10] The *n* number of measurements is grouped into regular intervals and the frequency within each interval has to be noted.

Figure 2.7 Frequency distribution of a random sample of measurements

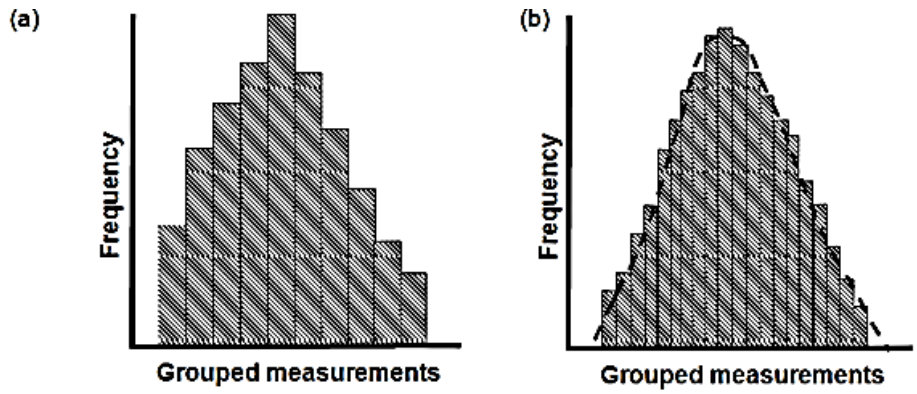
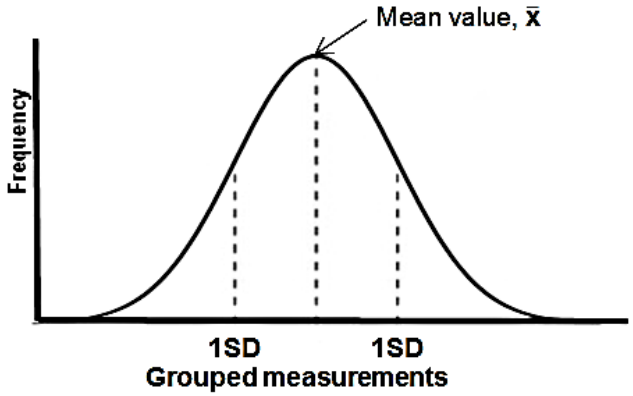


Figure 2.8 Distribution of repeated measurements



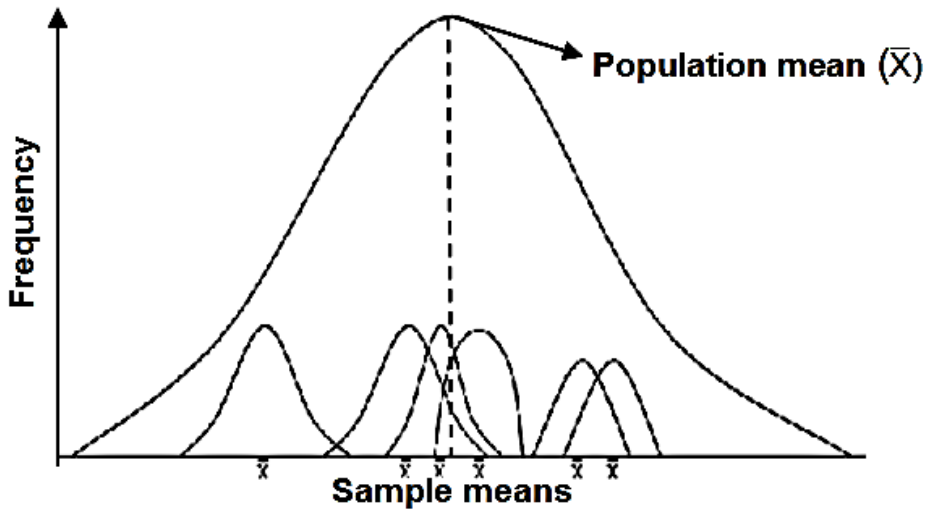
The area under the curve between any two measured values will give the number of measurements in that range of values. The curve is symmetrical around the mean value ($\bar{x}$), which lies in the middle of the distribution, and is the best estimate of the true value. Figure 2.8 shows the positions of ±1SD from the mean value. The SD characterises the spread of the repeated measurements around the mean; the bigger the SD, the greater will be the spread, which means the set of measurements is imprecise and the random errors quite significant. To reduce the SD, we can use a more precise measuring instrument or reliable measuring technique or better control of the variables. Since repeated measurements obey normal distribution, it means 68% or about two-thirds of measurements fall within 1SD from the mean value in both halves of the curve (see Figure 2.8). If the range is extended to 2SDs, the total area under the curve will represent about 95% of all measurements,

and this can be translated to mean 19 out of every 20 readings fall within the two measured values at the ±2SD marks. The preceding discussion implies SD can help predict the subsequent measurements of a particular quantity.

Under normal circumstances, repeated measurements are expected to stay within a normal distribution, but sometimes there may be anomalous (also known as “outliers”, “aberrant” or “abnormal”) results. Such a result has to be closely examined to determine its possible causes. If it is due to gross mistake or poor measurement procedure then the abnormal datum is discarded, but if it is part of the variation in a data set, it should be kept. If the anomaly is suspected to be caused by an uncontrolled variable, the investigator has to make further checks on the working conditions surrounding the investigation.

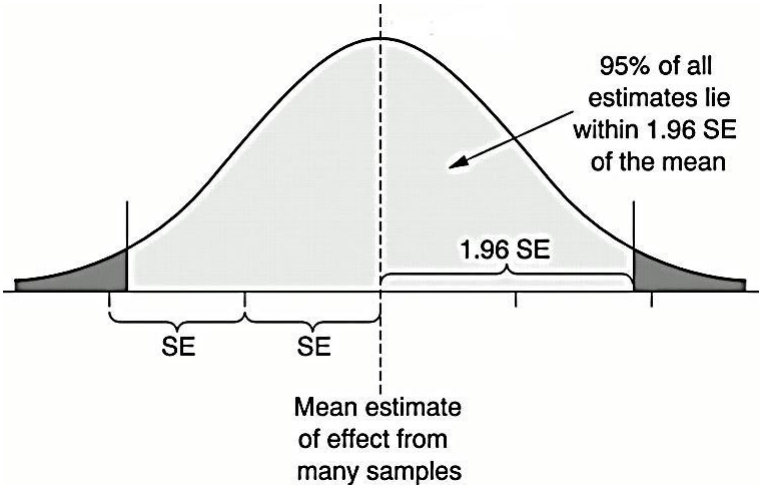
The SD, however, only describes the distribution of one random sample of measurements. The ideal will be to describe the whole population of measurements, which we can imagine as making up of all the different sets of data taken independently and each has a mean value that differs from the other by a small quantity. To describe the whole population, we can find the SD of the mean values of all these independent samples of measurements, which is known as the “Standard error (SE)” (See Figure 2.9).

Figure 2.9 Standard Error: the SD of sample means (image modified from http://www.ilri.org)



The SE can be estimated from a normal distribution of sample means by plotting a frequency distribution of all the mean values ($\bar{x}$) of different random samples of measurements (see Figure 2.10). SE, like SD, becomes a measure of the dispersion of sample means ($\bar{X}$) for a whole population of measurements.

Figure 2.10 Normal distribution of sample means (image taken from http://antongerdelan.net)



We can also estimate the SE from just one random sample of the entire population using Equation (4):

$$SE = \frac{SD}{\sqrt{n}} \quad (4)$$

Where SE = standard error
SD = standard deviation of a sample
n = number of measurements

The SE value can then become the measure of variability that exists between the mean value of a random sample and the true mean of a population. We can do this by using SE as *confidence limits* (for e.g., ±1SE or 68%, or ±2SE or 95%). To illustrate, suppose in the bouncing ball investigation, the mean value of 100 rebound heights is 41.0cm and the SD is 4.0cm, applying Equation (4), the SE is equal to 0.4cm. Therefore, our best guess of the mean rebound height for the whole population is 41.0±0.4cm. This means we are 68%[11] certain the true population mean lies between 40.6 and 41.4cm.

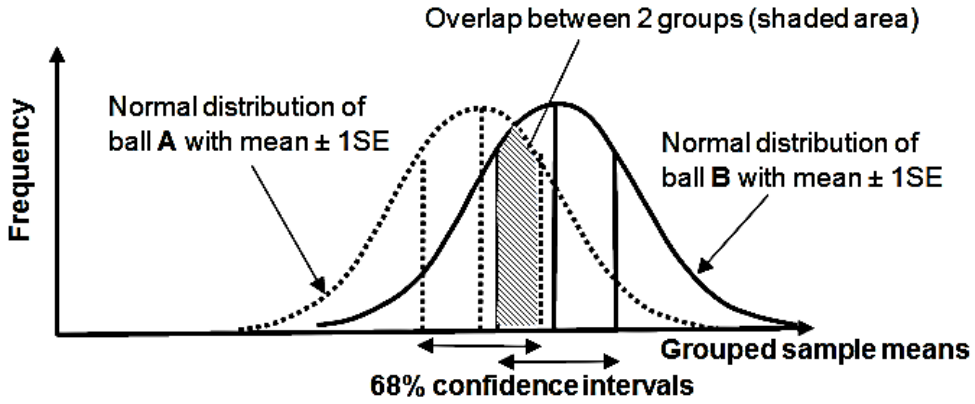
[11] Similar calculations can be performed to determine for 2SE (95%) and 3SE (99%).

As mentioned earlier, the SE enables us to “generalise about the population as a whole from one sample” (Gott & Duggan, 2003, p.151). In order to do this, we strive to make the SE small so that the random sample comes closer to representing the whole population based on its mean value. From Equation (4), there are two ways to achieve this.

First, we can reduce the SD by taking the best practical actions such as using a reliable instrument and deploying the proper method of measurement. While such actions may result in better precision, there is a limit to what can be achieved due to uncertainties caused by other factors which cannot always be experimentally controlled such as the effects of a fluctuating environment.

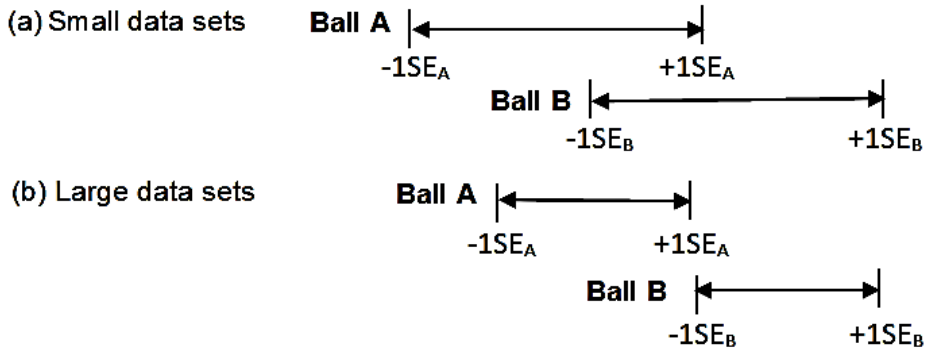
Second, since SE is inversely proportional to the square root of the number of measurements (*n*); we can reduce the SE by increasing factor *n* by a large margin. In fact, in most investigations, it is the second method that is often suggested; so, we carry out many measurements in order to get a smaller SE, and thereby bringing the sample mean closer to the true population mean value. This gives rise to the rule of thumb - the more repetitions we make of a measurement, the better the estimate will be (Guare, 1991). The idea of increasing the number of measurements to reduce the SE is especially useful if we want to establish the difference between two normal populations of a particular variable. To illustrate, we look at the bouncing ball investigation again. If we take small data sets of rebound heights of ball **A** and **B** to establish that ball **B** rebound higher than ball **A**, we may end up with an overlapping region since both balls sometimes rebounded to similar heights (see Figure 2.11).

Figure 2.11 Overlapping distributions of small data sets based on SE values



If we increased the number of measurements for each ball, we can reduce their SE values to the extent the distributions of both groups separate out clearly, and there is no overlap between them (see Figure 2.12). This will make us 68% confident the two balls have different bounciness and ball **B** normally bounce higher than ball **A**.

Figure 2.12 Effects of small versus large data sets on 68% confidence intervals



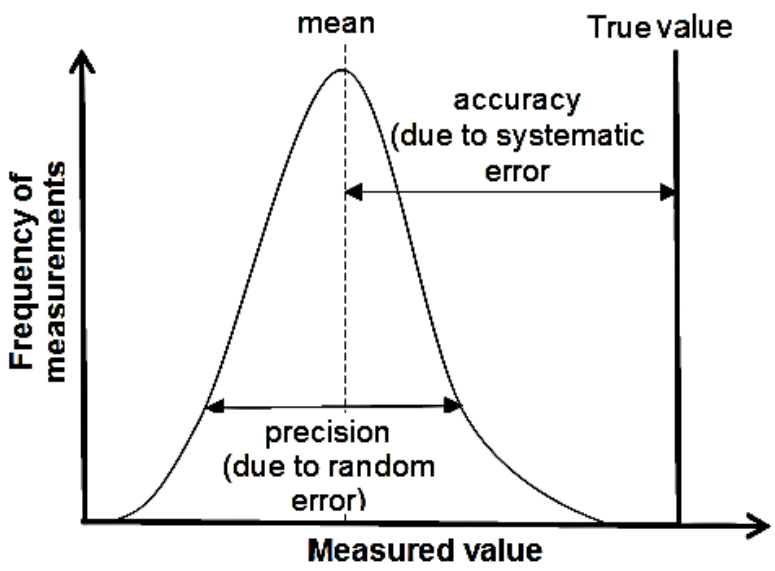
The SE can be used to plan the number of repeats (after trial runs to assess the variation in repeated readings have been conducted). From Equation (4), if we take ten instead of one measurement, it will give us improvements by reducing the SE by about a factor of three, which may seem quite attractive in many cases. However, the square root for the factor *n* being the denominator gives us diminishing returns, thus, if we want improvements by another factor of three, we may have to repeat the measurements a hundred times, which may not be feasible in a primary science investigation in terms of

time and logistical requirements. Nevertheless, for the PSTs, knowing the SE (say from a pilot) is useful as it assists them to balance between the needs for confidence in the data and the cost of collecting a large amount of it (Gott & Duggan, 2003).

2.2.7 An Overview of Uncertainty in Measurements

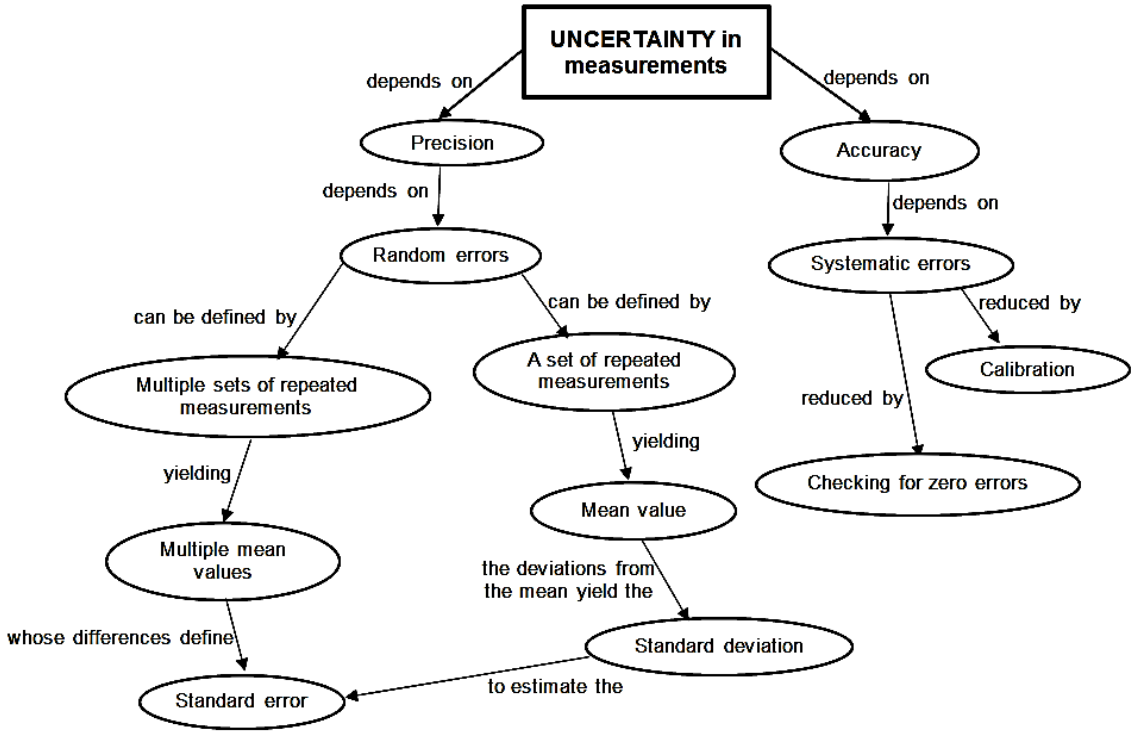
Distilling from the preceding discussions, we can see underpinning understanding of uncertainty in measurements are the concepts of accuracy and precision. These concepts are, in turn, dependent on experimental errors. Measurements can only be accurate if they are relatively free of systematic errors and precise if the random errors are reduced (see Figure 2.13).

Figure 2.13 The relationship between accuracy, precision, and experimental errors (modified from Heinicke & Heering, 2013)



The concept map in Figure 2.14 provides an overview of how uncertainty in measurements can be understood.

Figure 2.14 A concept map for understanding uncertainty in measurements



Section 1.6.3 described what constitutes “understanding”. Figure 2.14 implies that understanding uncertainty in measurements requires the ability to apply and synthesise key concepts/ideas like “accuracy”, “”precision” and “experimental errors”, which in turn requires the understanding of other concepts/ideas (some are not shown in Figure 2.14, for e.g. the resolution of scale, instrument reliability, uncontrolled variables, etc.) that may also be interrelated. It is the understanding of all these underpinning ideas of evidence that this thesis seeks to investigate in the PSTs.

The natural question to ask at this juncture is how this current research intends to go about exploring the PSTs’ understanding of uncertainty in measurements. This will be the focus of the next section.

2.3 Eliciting evidence for PSTs' understanding

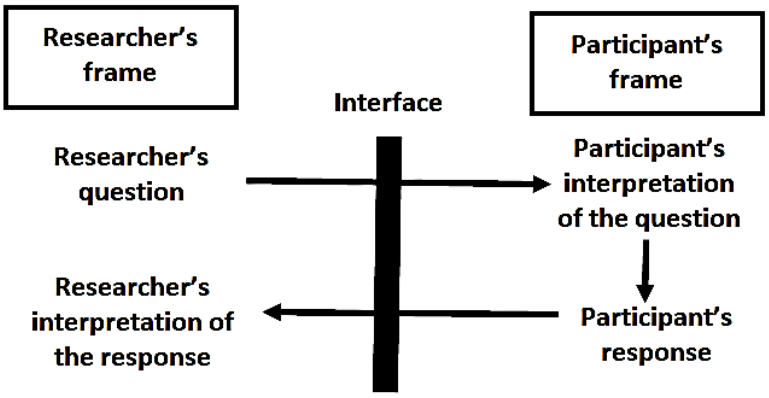
An important aspect of this current research is to determine what counts as evidence that a PST has a certain understanding of ideas of evidence. Several researchers have questioned the empirical basis of evidence in which students' conceptions have been founded (see for e.g., Lythcott & Duschl, 1990; Jones & Gott, 1998; Millar, 1998). Millar, for example, critiqued the methodology of several scientific reasoning studies for portraying scientific investigations as having an "invariant template"; he also claimed the research tasks were often "poor analogies for real situations where scientific thinking is required" (p.2).

Johnson and Gott (1996) too questioned the validity and reliability of some studies for assuming their interpretations of the participant's responses to questions they asked actually described what a participant might be thinking. Johnson and Gott argued this cannot be fully justified from a constructivist perspective as every individual (including the researcher) would be making their own meaning of the world (Johnson and Gott referred to this as the "frame of reference").

As a result of the differences in the frames of references for both researcher and participant, there would always be an "interpretation interface"[12] in the communication between the researcher and the participant. This interface has to be traversed twice, once when the participant interprets the researcher's question, and a second time, when the researcher interprets the participant's response (see Figure 2.15). At both times, interpretation differences can arise.

[12] Johnson and Gott (1996) originally used the word "translation" but now feel interpretation would be a better word to convey their real intent (personal communication).

Figure 2.15 The Interpretation Interface (modified from Johnson & Gott, 1996, p.564)



To reduce these differences, Johnson and Gott suggested setting up a "neutral ground" which they claimed is the "undistorted communication" (p.565) between the researcher and the participant. Herein, the aim is to allow the participant to understand what the researcher is asking in the meaning intended by the researcher, and the researcher to understand the participant's response in the meaning intended by the participant. It is important to note the use of the word "neutral" instead of "common" as the latter will be "precluded by the fundamental constructivist principle" (p.565). Johnson and Gott identified three basic components of the "neutral ground" that can form the methodological principles researchers can refer to in designing their studies and these are:

(a) "Neutral tasks": The tasks including the associated questions have to be *neutral* in two areas; firstly, in terms of being accessible to both researcher and participant, and secondly, the tasks should not constrain the thinking and possible responses for both parties. Accessibility can be with respect to the ability of the researcher or the participant to understand the question, for instance, an individual who is well-grounded in physics may have difficulty understanding a question that uses a biology task because it is an unfamiliar context. On the second point, thinking can be constrained if the questions appeared to be ambiguous, filled with technical jargons, or contained diagrams that were too complicated. All these might distract the participant from

answering the questions sufficiently resulting in the researcher not being able to fully elicit the participant's understanding. Johnson and Gott proposed ways to improve task neutrality, for example, by conducting pilot studies, or by providing opportunities for the researcher to check the participant's "real" understanding of the questions through consultation meetings or focus group discussions during or before the start of data collection.

(b) "Interpretation on neutral ground": The researcher must guard against imposing his or her frame of reference on the participant's responses. Instead, the researcher must attempt to understand the participant's answers based on the participant's own frame of reference. Thus, if a researcher is collecting data via interview, the researcher should check whether he or she has interpreted the participant's response accurately by paraphrasing the responses and by asking the interviewee several times in different ways. The researcher can also check the responses with other researchers to see if there is any disagreement or misinterpretation. The researcher can also clarify with the participant to see if the response data have been interpreted correctly and matched with the meaning intended by the participant.

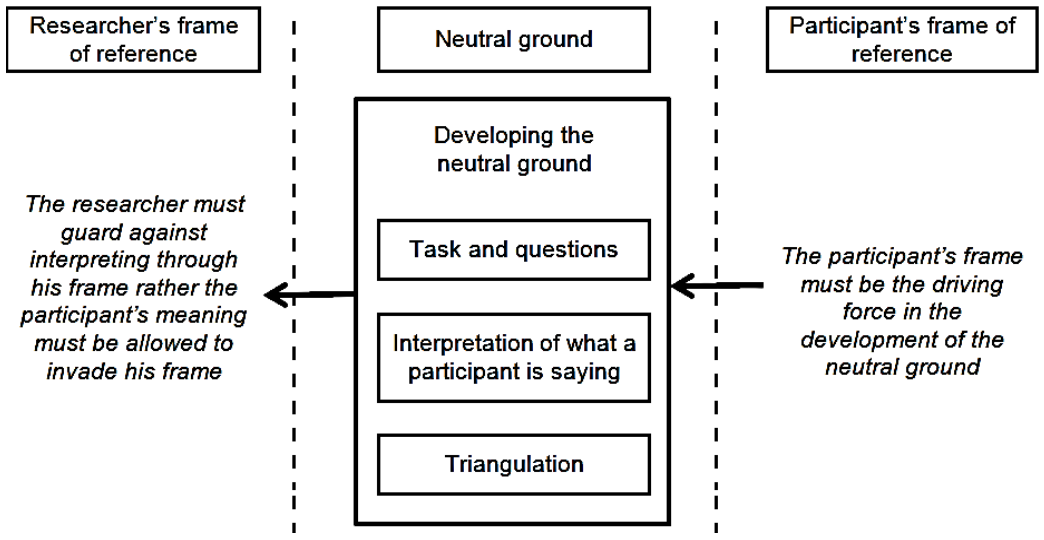
(c) "Triangulation": Even if (a) and (b) were to be carried out meticulously, participants might sometimes provide answers that did not truly represent their thinking or what they believed in. McClelland (1984) referred to this as "instant invention" whereas Piaget (1929; cited in Johnson & Gott, 1996) called it a "romancing response". In these cases, the validity of the participants' answers might be questioned on grounds that the responses were not reliable. To address this problem, Johnson and Gott suggested triangulation should be used and deployed in two ways: testing the same idea in a range of tasks that are related conceptually; and testing the idea at different times.

Other researchers have also suggested to use triangulation in qualitative research to help validate research findings (for e.g., Denzin, 1978; Jick, 1983), and to confirm research data (Miles & Huberman, 2002). Jick argued that when more than one method is used in the research process, the findings will help to cross-validate the "variance" (p.136) in the data and we can safely say the observations are simply not methodological artefacts peculiar to one research instrument. Jick also cited Denzin (1978) who suggested using "between-methods" and "within-methods" (p.136) to address the qualitative validity and internal reliability (consistency) respectively in a qualitative study. "Between-methods" can involve two or more independent but different modes of data collection (for e.g., interviews and questionnaires) to yield comparable data; if the data are congruent, then it will enhance the validity of the evidence. As for "within-methods", this can be done by deploying different ways of asking about the same concept within an instrument.

Jick also claimed that triangulation will result in a more holistic portrayal of the subjects as it provides the "thick description" (Geertz, 1973) a qualitative research seeks to have. In addition, the research will gain from the inherent strength triangulation offers as "the weaknesses in each single method will be compensated by counter-balancing strengths of another" (Jick, 1983, p.138). Finally, triangulation allows the checking of information from the participants and addresses the "interpretive validity" (Miles & Huberman, 2002) or getting an accurate meaning of the participants' responses in a study (concurring with arguments forwarded by Johnson and Gott earlier).

Figure 2.16 below gives an overview of how the "neutral ground" can be developed. This research intends to follow such an approach in pursuit of its goal of exploring the PSTs' understanding of uncertainty in measurements.

Figure 2.16 Developing the "neutral ground" (modified from Johnson & Gott, 1996, p.568)



In order to develop the "neutral ground", we can first learn how other researchers have attempted to investigate their participants' understanding of uncertainty in measurements. This is the intention of the next section.

2.4 Review of studies on understanding of uncertainty in measurements

The term "procedural understanding" was introduced in Section 1.6.3 in Chapter 1. According to Newton (2000), procedural understanding is one of several basic kinds of understanding that we should aim to develop in science education alongside conceptual (understanding of scientific laws, theories and concepts like energy and photosynthesis), situational (understanding what a situation amounts to and describe it), and causal (understanding how one thing leads to another, and the cause-and-effect relationships) understandings (see reference for a fuller description). All four are important considerations for primary science teachers as they design activities, including science investigations, to promote deep understanding of science in their lessons (Newton, 2001).

Beyond the development of deep understanding of science, the current trend in science education (including in Singapore; see Chapter 1) is also

moving towards the incorporation of “scientific literacy” into the school science curriculum (American Advancement for the Advancement of Science [AAAS], 1993; Hurd, 1998; Laugksch, 2000; MOE, 2008; Organisation of Economic Co-operation and Development [OECD, 2013). Such endeavour is often driven by a national objective such as: “citizens …should understand how science works and how it is based on the analysis and interpretation of evidence” (UK House of Commons Science and Technology Committee, 2002; p.36). Since “procedural understanding” is important in developing scientific reasoning skills (Newton, 2001) and the “thinking behind doing” of scientific investigations (Gott et al., 2008), naturally, many seem to see “procedural understanding” as being an essential part of scientific literacy education (Millar & Osborne, 1998; Ryder, 2001; Bybee, Powell, & Trowbridge, 2008). Additionally, many may view it as an integral part of developing ideas about nature of science (Abd-El-Khalick, Bell, & Lederman, 1998; McComas & Olson, 1998; Murcia & Schibeci, 1999; Abd-El-Khalick et al., 2004) whereby a large body of research has been conducted on investigating students’ epistemological understanding of the nature of evidence, for instance, how students’ view the images of science (for e.g., Ryder & Leach, 1999; Séré et al., 2001), how scientists work (for e.g., Petkova & Boyadjieva, 1994; Hogan & Maglienti, 2001), the relationship between theory and evidence (for e.g., Leach, 1999; Ryder, 2002; Guerra-Ramos, Ryder, & Leach, 2010), and the construction of knowledge through inquiry (for e.g., Lin, Chiu, & Chou, 2004; Sandoval, 2004, 2005; Wu & Wu, 2011). These studies have largely shown that students’ epistemological ideas may influence their handling of evidence.

Nevertheless, the review done here will not cover studies in this area because of the limitations of time and word count. The research done here

instead intends to explore more of the “what” question - the idea-base of evidence that the PSTs have with respect to handling uncertainty in measurements. Borrowing from Séré (2002), the objective will be to identify “what remains conscious… and how an awareness of [ideas] helps [the PSTs] to decide, plan, design, and realize experiments on their own” (p.627).

To my knowledge, there is a lack of research that specifically looked at *pre-service teachers'* understanding of uncertainty in measurements, and the extent in which such understanding can influence their decisions during investigations or classroom instructions. It is indeed ironical that PSTs are trained to provide students with inquiry experience aimed at developing procedural understanding including handling uncertainty in measurements and yet not much is known about how the PSTs themselves understand the concept. Nevertheless, since teachers were once students, it would be reasonable to review the literature with the intention of drawing learning points from it. Table 2.1 provides the outcomes of this review where key findings relevant to this research are stated; these will form the basis in which the findings of the current research can be compared and contrasted.

As a whole, the body of literature that looks at uncertainty in measurement is rather “unbalanced”. There are not many quantitative studies on understanding uncertainty in measurements; only a few studies were carried out on a large scale. Most were one-off qualitative studies that employed convenience sampling and self-reporting methodologies such as interviews and questionnaires. This is not to say that qualitative methods were inferior in any way, rather it reflects the difficulties incurred in attempting to quantify a complex variable such as understanding uncertainty in measurements. Additionally, most studies were based on the physics context, and very few on other science

subjects. Finally, the range of studies represented a broad spectrum of educational systems attending to different objectives, and this rendered comparison of research findings difficult.

Table 2.1 Summary of empirical studies on uncertainty in measurements

Study	Objectives	Research Design/ Methods	Description of sample	Key Findings
1. Séré, Journeaux, and Larcher (1993)	To determine students' application of concepts of measurement (for e.g., true value , accuracy, precision) and statistical concepts (for e.g., estimation of errors) after a theory course	Laboratory report on finding focal length of lens; Final written test on electricity; Post-test interviews with selected students	20 first-year French physics undergraduates; 18-20 years	The students established a routine to deal with laboratory measurements but did not really understood the purpose of repeats and why "the more measurements, the better"; they had poor understanding of how to apply their statistical knowledge to measurements, and had difficulty distinguishing conceptual differences between terms like accuracy and precision.
2. Millar, Lubben, Gott, and Duggan (1994)	To determine children's ability to carry out science investigation tasks, and the understandings that inform their actions (Procedural and Conceptual Knowledge in Science or PACKS Project)	Observations and students' recordings of one of 7 investigative tasks (different contexts), followed by interviews, as well as written diagnostic probes	800 UK primary and lower secondary students at three age points between 9 and 14 years(Year 4,6 or 7, 9)	Children's understandings of the aims and purposes of investigating, and ideas of evidence underpinning criteria for evaluating the quality of empirical data were important factors in determining children's performance of an investigative task (alongside their understanding of substantive science concepts relevant to that specific task).
3. Lubben and Millar (1996)	To explore students' purpose for repeats, ways of handling repeats, anomalous measurements, and the significance of spread in results	2 questionnaires, each consisting 6 different probes	1040 UK primary and lower secondary students, 400 Year 7(11 years), 400 Year 9(13 years), and 240 Year 11(15 years) students	Identified a pattern of progression with age and experience in understanding the need to repeat measurement, the evaluation of measurements, and handling of anomalous results; students described a fair test as one that had equal number of repeats in the data sets being compared.
4. Varelas (1997)	To explore the sense of variability and the idea of a representative of repeats	Videotapes and transcripts of recordings of group investigations; field notes	24 Year 3 and 4 American students; 8-10 years	Students in the early age groups were not able to conceptualise the procedure of repeating trials and finding the best representative of repeats.

Study	Objectives	Research Design/ Methods	Description of sample	Key Findings
5. Allie, Buffler, Lubben and Campbell (1998)	To determine students' understanding of measurement during data collection, processing and comparison; how to apply the model of progression (Lubben & Millar, 1996)	Questionnaire consisting of 6 probes based on a single experiment (ball rolling down a ramp and landing on the floor); questionnaire administered prior to giving instructions on the subject; interviews conducted to validate observations with selected students	121 first-year South African physics undergraduates in a foundation programme; 18-21 years	Most students repeated readings to confirm a reading, obtain a true value or a mean value; many did not understand the mean together with the spread represented the quality of data and not the mean value alone; most students appeared to be "spread reasoners". The researchers proposed students could be classified as "point" and "set" reasoners.
6. Coelho and Séré (1998)	To determine students' difficulties in reasoning during data collection, processing and interpretation and to explore students' concepts of true values	Clinical interviews during investigation on finding the constant velocity of an air-puck	21 French secondary students; 14-17 years	Most students believed true values existed along with a "perfect" investigator or "perfect" instrument. Students favoured mode and median to represent their repeated measurements other than the mean values.
7. Leach et al., (1998)	To explore students' views about uncertainties and sources of uncertainty, the differences between accuracy and precision, and how they overcome uncertainties and select a value	Questionnaire comprised of 3 probes accompanied by questions to give supporting reasons; based on the context of nutritionists trying to measure the mass of oil samples	422 upper secondary and 229 science undergraduates from 5 countries: Denmark. France. Germany, Great Britain, and Greece	Students tended to think it was possible to make a "perfect" measurement; 5% thought their measured data were true values; 30% of students felt the arithmetic mean was enough to compare two sets with same mean value but different spreads; in most probes, the undergraduates seemed to be able to give responses closer to the intended responses compared to the other age groups showing such understanding might be age-related.
8. Millar (1999)	To study the differences in students' actions when investigating the effect of an IV which does not alter the DV, and when investigating the effect of an IV which does	2 computer- simulated investigations of a pendulum experiment	30 English students age 14	Students took significantly more measurements when investigating the effect of IV (which produces no effect), and there were also significantly more instances of repeating a measurement. Students seemed to have difficulty in reaching a conclusion about the effects of IV when the DV measurements had uncertainty.

Study	Objectives	Research Design/ Methods	Description of sample	Key Findings
9. Evangelinos, Psillos and Valassiades (1999)	To investigate students' views about the nature of single measurements	Written questionnaire of 6 different probes; questionnaire administered prior to laboratory instructions	32 first-year Greek physics undergraduates; 17-18 years	Students had deeply-embedded views about the exactness of measurements, and found difficulties in accepting uncertainty. Students generally viewed a measurement as either exact, an approximation of the true value, or an interval.
10. Buffler, Allie, Lubben, and Campbell (2001)	To determine use of "point/set paradigms" at the end of a laboratory course	Same as 5. Intervention study with a pre- and post-test design; 1 probe on data comparison was added in the pre-test, and 1 on data collection was removed in the post-test because of duplication	147 first-year South African physics undergraduates in foundation programme in the pre-test; 125 remained in the post-test; 18-21 years	Significant shift from being consistently "point" to "set" reasoners; majority represented a data set by mean values but still failed to consider spread; lots of confusion over terminologies like spread, range, etc.
11. Lubben, Buffler, Allie, and Campbell (2001)	To determine the usefulness of "point/set paradigms" for interpreting ideas about measurement and uncertainty	Same as 5. 7 probes including 2 new probes on data comparison and the removal of 1 probe on data processing	257 first-year South African physics undergraduates (174 in foundation programme);18-21 years	Most students were consistently using either "point or set paradigm". Measurement decisions might be dependent on the task context.
12. Rollnick, Dlamini, Lotz, and Lubben (2001)	To examine reasons for repeats and ideas of handling data in a chemistry-based context	Questionnaire with 7 probes adapted from Allie et.al. (1998) and based on a single task of a precipitation reaction. Administered prior to giving instructions	231 South African chemistry undergraduates in a "bridging programme" at the start of university; 18-21 years	Most repeated to get recurring reading, practice or to get a mean value. Evidence showed the mean value calculation became a procedural routine and there was poor understanding of spread. Most believed anomalous readings should not be excluded in mean value calculations; and had difficulties distinguishing data sets with the same mean value but with different spreads.

Study	Objectives	Research Design/ Methods	Description of sample	Key Findings
13. Ryder and Clarke (2001)	To determine about sources of error in science, and the impact of explicitly teaching the understanding of errors	Questionnaire followed by interview to validate response at the beginning and the end of a physics module. Questionnaire was based on the context of a digital thermometer	7 English second year undergraduates attending a physics module; 18-20 years	Students needed to have qualitative reasoning for the sources of systematic/random errors; know how to distinguish between systematic and random errors and their relationship to the quality of measurements; and how to apply conceptual understanding of errors to different measurement contexts.
14. Tomlinson, Dyson, and Garratt (2001)	To explore understanding of key terms and concepts used in dealing with experimental errors and uncertainty	Questionnaire with 5 sets of open-ended questions on 12 common terms	103 first-year English chemistry undergraduates;18-21 years	Little or no understanding with terms such as precision, accuracy, and random errors; evidence showed students were more inclined to link variation in measurements with mistakes than with random errors.
15. Evangelinos, Psillos and Valassiades (2002)	To study students' concepts and reasoning of instrument readings by comparing 2 groups who received "innovative" and "conventional" instructions	Intervention study with a pre- and post-test design; multiple-choice questions followed by justification	First-year Greek physics undergraduates; 16 in "innovative" group and 51 in "conventional" group;17-18 years	"Conventional" group, unlike the "innovative" group, viewed measured values as approximates of the true value and tended to ignore their probabilistic nature.
16. Leach (2002)	To determine the sources of error in calculated kinetic values, and why data did not agree with predicted theoretical values	Survey then interview of selected students. Study was contextualised in measuring enzyme kinetic data	48 first-year English biochemistry undergraduates, 6 were selected for interview; 17-19 years	Students believed true values were measurable, errors were mistakes, repeats gave insights into accuracy rather than precision; and one could judge an estimate from a data book and not from the quality of the data set it was estimated from.

Study	Objectives	Research Design/ Methods	Description of sample	Key Findings
17. Buffler, Allie, Lubben, and Campbell (2003)	To evaluate new course based on the probabilistic framework of measurement	Same as 5. Intervention study with a pre- and post-test design; 1 probe on single measurement and 2 probes on the concept of true values were added to the questionnaire	106 first-year South African physics undergraduates in foundation programme; 18-21 years	Students who chose “approximate” in the pre-test became more “set” reasoners in the post-test; there were less students who claimed the existence of true values and “perfect” measurements after the course.
18. Lubben, Campbell Buffler, and Allie (2004)	To determine how measurement context influence students ’ perceptions of measurement quality	Questionnaire comprised of 7 probes; 3 probes on different weighing situations and 3 probes on analogue readings, and 1 on digital display	78 South African physics first-year undergraduates; 18-21 years	Three types of judgements involved: results -driven, process-driven, and instruction-driven. Majority seen to use one judgement criterion consistently across different contexts; thus, the interpretation of measurements depended on the individual’s fundamental criterion for judging measurements.
19. Kanari and Millar (2004)	To explore reasoning in two investigative tasks which consisted of 2 IVs and a DV where one IV co-varied with the DV but the other did not	Interviews based on video-recordings of 2 tasks: investigation of the time of swing with length and weight of pendulum, and relationship between mass of small box and area of its bottom surface and the pulling force	60 English students of 10, 12 and 14 years of age(20 each)	Students had much higher success rate at concluding from an investigation where the DV co-varied with the IV compared to one in which the DV did not vary with the IV. In data collection, students were more “trend-focused” than “difference-focused” (one or more pairs of values of the IV such as a high and a low value) to see if this resulted in a difference in the value of the DV).
20. Kung and Linder (2006)	To determine how students would report experimental data, compare data quality based on the number of trials, and how they combine two sets of data and represent the combined set	Intervention study with pre-and post-survey; Questionnaire comprised of 7 open-ended questions based on the same task of a ball rolling down a ramp used by Allie e.al.(1998)	41 second-year Swedish engineering undergraduate;19 -21 years	Post-survey showed more included spread to report uncertainty in results; reported uncertainty when they combined their data from two sets of repeats; were able to compare two sets of data with the same mean but one had a larger spread; suggested more trials other than a fixed number of three. Overall, students seemed to have moved from the idea that “smaller range means better results” to “better result means better known average”.

Study	Objectives	Research Design/ Methods	Description of sample	Key Findings
21. Heisawn, Songer, and Lee (2007)	To determine students' ability to collect and interpret evidence based on the notion of "evidentiary competence" (concepts and reasoning skills involved in collecting, organising, and interpreting data)	12 questions; mixture of open-ended and multiple-choice and justification format contextualised in the topic of atmospheric science	40 American Year 6 students; 11-12 years	Most students did not understand inherent variability and did not repeat because they did not understand that repeats were needed to "compensate" for uncertainties; less than half of students given a tabulated data (with uncertainties) were able to articulate the relationship between the two variables involved.
22. Åkerlind, McKenzie, and Lupton (2011)	Having identified uncertainty in measurement as a Threshold Concept (Meyer and Land, 2005), the study aimed at investigating the critical aspects that must be understood to fully grasp its meaning, and get over the "threshold"	Interviews using two probes followed by phenomenographic analysis of transcripts	23 Australian first-year physics undergraduates;18-20 years	Three critical aspects to understanding uncertainty: a pattern-recognition element that allowed students to distinguish between trends, noise, and anomalies; a formal procedural understanding that allowed quantifying and combining different elements of uncertainty; and a "meaning" element that invested uncertainty with meaning that had implications beyond the given data. A sophisticated understanding involved the integration of all three aspects.
23. Munier, Merle and Brehelin (2012)	To identify and study the development of reasoning for measurement variability.	Longitudinal study over a year. Field notes, video-recordings, written tests, and clinical interviews. Used the contexts of measuring instruments (calliper and digital balance)	24 French Year 4 students(9-10 years); and 22 remaining students(Year 5;10-11 years) one year later	After instructions, students appreciated the relationship between the quality of instrument and precision; students understood that spread could be due to instrument, the investigator, and the object being measured; their relative contributions depended on the measurement situation.

From Table 2.1, we could draw some common observations about the understandings of uncertainty in measurements:

- Many students might not fully grasp the meanings of concepts related to uncertainty or they cannot distinguish between the various concepts. Some notable examples included “error” (Leach, 2002), “accuracy” and “precision” (Séré et al., 1993; Evangelinos et al., 2002); “reproducibility” and “repeatability” (Tomlinson et al., 2001). Because of their poor comprehension, students generally have difficulties articulating their understandings including applying the concepts. There were also cases of misconceptions notably errors were thought of as “mistakes” (Leach, 2002), variations in data were caused by “mistakes” instead of random errors (Tomlinson et. al, 2001), a “fair test” involved a comparison between two data sets with equal number of repeats (Lubben & Millar, 1996).
- In handling uncertainties in single measurements, the concerns were mainly with the characteristics of a measuring instrument (for e.g., from Sere et al.(1993): the students were concerned the ruler they used did not have a small enough resolution of scale to measure distance accurately), and the way the measurements were carried out (for e.g., in Coelho and Séré (1998): difficulty in interpolating between two divisions on a ruler), and whether students view measurements as a probabilistic value (Evangelinos et al., 1999, 2002).
- Students did not fully appreciate the contribution of systematic/random errors towards uncertainty (for e.g., Séré et al., 1993, Ryder & Clarke, 2001), and attributing uncertainty mainly to human or instrumental errors which they felt could be eliminated by deploying a “perfect”

instrument/procedure(for e.g., Séré et al., 1993, Coelho & Séré, 1998, Leach et al., 1998).

- Separate studies consistently showed students believing in the existence of true values that should be unaffected by experimental errors and did not contain uncertainty. Students believed they could obtain such “perfect” values in their measurements (for e.g., Séré et al., 1993; Coelho & Séré, 1998; Leach et al., 1998; Evangelinos et al., 1999).
- Students especially those in the early years might not understand the purpose of repeating measurements (for e.g., Varelas, 1997; Lubben & Millar, 1996; Heisawn et al., 2007). Students at a younger age might also have difficulties handling anomalous results and evaluating repeats because the ideas could be too complex; understanding appeared to progress with cognitive development (Lubben & Millar, 1996).
- Variation in repeated measurements was the focus of many studies (for e.g., Varelas, 1997; Kung & Linder, 2006; Heisawn et al., 2007). One difficulty for many children (and some adults) was in distinguishing the variation due to random errors and the differences arising from changing the magnitude of an independent variable. This could lead to the inability of fully understanding the relationships between the variables in the investigation (for e.g., Millar 1999; Kanari & Millar, 2004).
- Many students appeared not to appreciate the significance of spread (or variation) in a data set when estimating a quantity since the mean value was all that matters (for e.g., Allie et al., 1998; Leach et al., 1998; Lubben et al., 2001; Kung & Linder, 2006). Students had difficulties

when they compared data sets if the spreads in each set overlapped or the mean values were the same.

- There were also problems associated with applying statistical concepts like the mean, SD (or confidence intervals), and SE (for e.g., Séré et al., 1993, Leach et al., 1998; Buffler et al., 2001, 2003; Kung & Linder, 2006). Importantly, concepts like the "mean value" and "confidence intervals" as well as ideas like "the more measurements, the better" were often treated as routine ideas without any real understanding of their purpose. Some ideas like "three repeats are enough" became so entrenched that they interfered with the reasoning and interpretation of data (Kung & Linder, 2006).

The points showed the students had many difficulties understanding uncertainty in measurements; these stemmed from not comprehending key concepts to not knowing the purpose of certain procedural ideas and the inability to apply and synthesise procedural concepts for judgements about the quality of measurements. The problems might be further compounded by a lack of statistical thinking and understanding in applying statistical concepts like the mean, SD, and SE, etc.

We could also see students with a fragmented knowledge of uncertainty who relied mainly on rote ideas to make sense of their data as well as those who did not see the inherent variability of measurements believing a datum could be a true value and uncertainties were merely products of poor measuring techniques or faults in the instrument.

In addition, many studies seemed to have investigated students' understandings of uncertainty in DV repeats belonging to a Type 1 investigation (with a categoric IV), much less in DV repeats in a Type 2 investigation (with a

continuous IV) or single measurements. Thus, less was known about how students understood the uncertainties in these measurements. The current research shall attempt to fill these gaps to a certain extent.

The next part of the review intends to report on the method of study including the use of different research instruments. However, it may be too voluminous to critically report on the studies listed in Table 2.1 in this thesis. It may be more profitable for this thesis to report on selected studies in greater depth[13] to draw on the learning points to inform its own research methodology. As seen in Table 2.2, seven studies covering a range of different instruments and types of measurement have been selected for this purpose.

Table 2.2 Studies selected for methodology review

Studies	Types of measurement	Main instrument
1. Evangelinos et al. (1999) 2. Evangelinos et al. (2002)	A single measurement or an instrument reading	Questionnaire
3. Allie et al. (1998) 4. Lubben et al. (2001) 5. Buffler et al. (2001)	Repeated measurements of a DV (time/distance measurements of a ball landing on the floor after rolling down from a fixed height of a ramp)	Questionnaire
6. Séré et al. (1993)	Single and repeated measurements for a DV in a laboratory investigation setting (focal length of lens)	Laboratory worksheet
7. Coelho and Séré (1998)	Single and repeated measurements for a DV in a laboratory investigation setting (constant velocity of an air puck)	Interview during investigations

The critical evaluation of these seven studies will be guided by the arguments provided by Johnson and Gott (1996) for developing a "neutral ground"; thus specifically, we shall look at the questions that were asked (for e.g., to check for ambiguity and possible misinterpretations), the methods and

[13] The seven studies were selected not only because they had often been cited in the literature but also because they revealed many details about their instruments and how their research data were interpreted. Most other studies did not have such details, thus making critical evaluation an unenviable task.

instruments used in eliciting their participants' response, and the way the response data had been interpreted by the researchers themselves.

The review begins with the study by Evangelinos et al. (1999, 2002) that claimed a single measurement should be interpreted as a "probabilistic" value. The researchers believed the traditional emphasis on experimental errors was unhelpful since this resulted in students adopting a more "deterministic" mindset towards measurements and treating readings as "exact" values. Their theoretical framework basically emphasised the notion of the "degree of belief" (Duerdoff, 2009) in a measurement after accounting all its uncertainties.

Evangelinos et al. collected research evidence via a written questionnaire consisting mainly multiple-choice questions followed by open-ended questions that sought justifications for the choices. They claimed their students were mostly using a reasoning scheme shown in Table 2.3.

Table 2.3 Students' reasoning scheme (modified from Evangelinos et al., 1999, 2002)

Exact	Has a naive view that a scientist can, in principle, obtain the "true value" using a high precision instrument. Physical quantities are conceived to be exact quantities like real numbers geometrically represented as points on an axis.
Approximation	Has a pragmatist's view that exact determination will never be feasible for practical reasons such as the measurements will be affected by the lack of precision in instruments, human error, and environmental conditions. Physical quantities are still represented as a unique numerical value, but could be a slight deviation from a central value.
Interval	The whole interval is seen as a single value and the true value can be one of many possible values within it. An interval is given either because the investigator is keen to project a "safe" reading leading to a guaranteed conclusion or to generate confidence in the results of an investigation whose experimental conditions were far from ideal and the data appeared to have deviated from the established values.

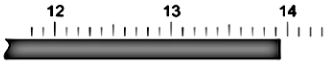
In their reports, the researchers provided two examples of multiple-choice question and both were concerned with the interpretation of an instrument reading (Figures 2.17 and 2.18). This review therefore focuses on these two

questions assuming they were good representatives of the rest in the instrument.

Figure 2.17 The interpretation of a measurement (Evangelinos et al., 1999)

A researcher used a high precision ruler to measure the dimensions of a body, as is depicted in the following figure. Which of the following sentences do you believe is best for reporting his measurement?

Please justify in each case.



The true length of the body:
(a) is exactly 13.93 cm, because ...
(b) is approximately 13.93 cm, because ...
(c) is between 13.9 and 13.95 cm, because ...
(d) is 13.95 ± 0.05 cm, because ...
(e) is other....., because ...

Figure 2.18 The interpretation of a measurement (Evangelinos et al., 2002)

An automated measurement set-up using a high precision chronometer gave a reading of 1.55834 sec. Which conclusion can a researcher draw from this reading and which not? Justify in each case.

(a) The true value of time is exactly 1.55834 sec.
(b) The true value is approximately 1.55834 sec.
(c) The true value is definitely lies between 1.558335 and 1.558345 sec.
(d) The true value probably lies between 1.558335 and 1.558345 sec.
(e) If we repeat this measurement, we can determine the true value of pressure with as many decimal figures as we please.

[*(e) was added and the chronometer was substituted by a manometer in the post-test*]

According to the researchers, those who chose (d) in both Figures 2.17 and 2.18 picked the best option, and could have a more probabilistic view of measurements (provided their reasons did not contradict their choices). Based on Table 2.3, the researchers described those who opted (a) as “exact reasoners”, (b) as “approximate reasoners”, and (c) as “interval reasoners”. Option (a) supported the idea of a true value and most students in both studies were able to see this. Although the numbers who chose (b) were slightly more than for (a), they were still quite low compared to (c); those who opted (b) might have thought of the measurement as being slightly off the actual value because of error(s), but their reasoning still affirmed the idea of a true value. Option (c) was the biggest distractor, and according to the researchers, it lacked the “probabilistic conclusion in terms of a confidence interval” (p.185), which was implied in option (d) by the presence of “±0.05” in Figure 2.17, and the term “probably” in Figure 2.18. Finally, the results for option (e) found in Figure 2.18

indicated the students acknowledged conducting repeats *ad infinitum* would never lead to a true value.

Looking at Figures 2.17 and 2.18, one could argue the questions did not test understanding; rather, they were testing the knowledge of terms that described uncertainty according to the "probabilistic" framework. Such a notion was implied in the researchers' comments of students' responses:

> …most students in their justifications *explicitly* name the interval as uncertainty and use the concept of probability, although these concepts were purposely not included in the wording of the task (Evangelinos et al., 2002, p.187).

Terms like "approximate", "between values" or "definitely between", "probably lies between" or "±" that appeared in the options might not be easily distinguished by students who were untrained about their technical differences; the expressions might have appeared similar to the PSTs during pre-test. Those who thought (b) and (d) were similar could be using the term "approximate" to mean "estimate"[14]. Thus, a finding like "the majority of students in the pre-test did not realise the four alternatives (a) to (d) were mutually exclusive, but agreed with both (b) and (d) providing similar arguments" (Evangelinos et al., 2002, p.188) was not unexpected.

Additionally, options (c) and (d) did not negate the "exact reasoners" who could be thinking in terms of discrete measurements with a limited rather than a continuous number of digits. In fact, the same argument could also be applied to those who opted (d) in the post-test; it remained inconclusive whether the students really understood the "probabilistic" nature of measurements or were merely recalling the routine method of reporting a measurement.

The term "high precision" used in both questions also seemed unnecessary, and could even be misleading as students might think the

[14] The term "estimate" is often used in place of "approximate" (Collins English Dictionary, 2014)

instruments presented had negligible errors prompting them to think the measurements should fall within the intervals of option (c) in both questions (i.e., “between” or “definitely lies between” values). Finally, the diagram of the scale provided in Figure 2.17 looked like a normal ruler only drawn bigger, and this could have resulted into thinking it was one (and not from a high precision ruler) and thereby caused misinterpretation.

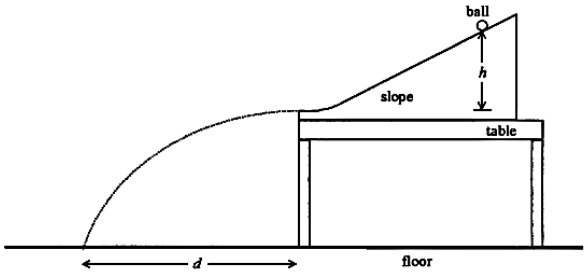
Next is the series of studies conducted in South Africa by Allie, Buffler, Lubben and Campbell between 1998 and 2003. Their investigations were largely based on uncertainties in repeated measurements. Their main instrument, a questionnaire, consisted of probes based on a single experimental task shown in Figure 2.19.

Figure 2.19 Experimental task used for the questionnaire (Campbell et al., 2005, p.14)

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***.



The reasons for using a single task, according to Allie et al., were to avoid confusion by the use of too many scenarios and to prevent cognitive overload for respondents who had little exposure to practical work. The probes were designed to look at students’ decisions during different phases of their investigation: collection, analysis, and interpretation of data (see Table 2.4).

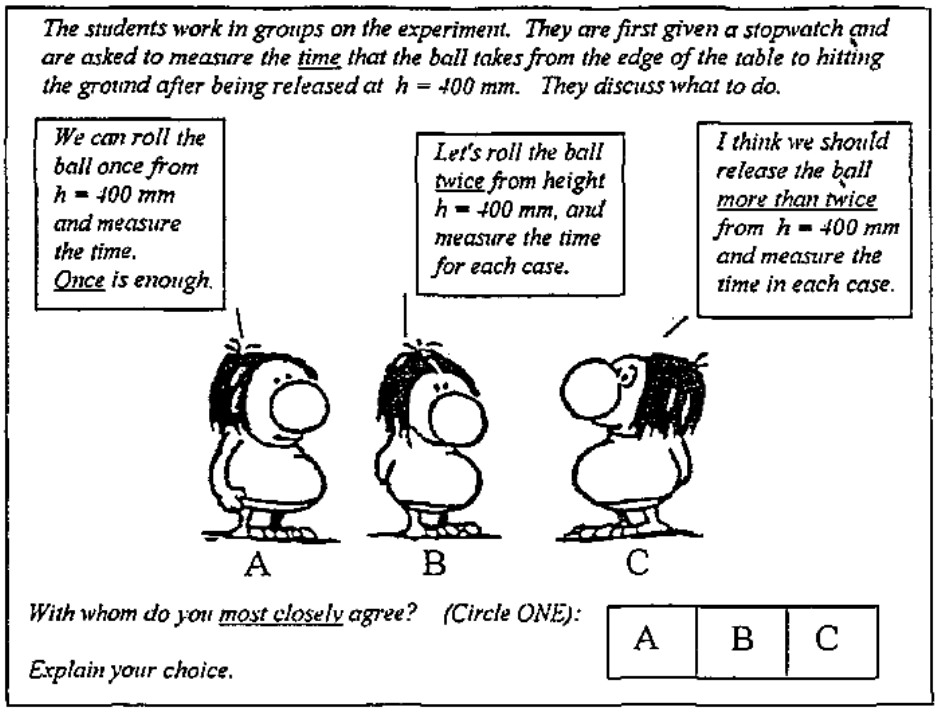
Table 2.4 The probes used in the South African studies (modified from Campbell et al., 2005, p.16)

Probe code	Name of Probe	Aspect of measurement	Used in
RD	Repeating Distance	Data collection	1,2,3,4
RDA	Repeating Distance Again		1,2,3
RT	Repeating Time		1,2,3
UR	Using Repeats	Data processing	2,3
AN	Anomaly		1
SLG	Straight Line Graph		2,3
SMDS	Same Mean Different Spread	Comparison of results	1,2,3
DMSS	Different Mean Similar Spread		1,2,3,4
DMOS	Different Mean Overlapping Spread		3
DMSU	Different Mean Same Uncertainty		3

Key: Study 1 – Allie et al. (1998); 2 – Lubben et al. (2001); 3 – Buffler et al. (2001); 4 - Buffler et al. (2003)

The probes were not developed all at once; rather, it started with six probes in the study by Allie et al. (1998), and subsequently more were added and some removed in the remaining three studies as the researchers examined new research questions (see last column in Table 2.4). An example of a stem from one of the probes is shown in Figure 2.20.

Figure 2.20 Repeating time (RT) probe (Allie et al., 1998, p. 450)



How were the responses to the probes analysed? In the first study, the responses to the probes were codified for common themes, and then analysed using Grounded Theory (Strauss & Corbin, 1990). A group of students was subsequently interviewed to check their understandings of the questions, and

the researchers' interpretation of the responses. The same method was claimed to have been deployed whenever new probes were created in subsequent studies.

The kinds of responses given by students in the first two studies were largely similar. Thus, when the response data from all the probes were analysed and interpreted, the researchers found their students' understandings of repeated measurements could be characterised by a certain pattern which they described as "point" or "set" reasoning (see Table 2.5).

Table 2.5 "Point" and "Set" reasoning (Campbell et al., 2005, p.30)

Point reasoning	Set reasoning
The measurement process allows determination of 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.

There appeared to be some difficulty, however, in teasing out the evidence according to the proposed reasoning paradigms; for instance, the researchers accepted the concept of mean value as indicative of a "set reasoner", but students who were preliminarily classified as "point reasoners" used mean value to respond to other probes later. There were far too many of such cases (including those that dealt with concepts other than the mean value) that made prediction using Table 2.5 difficult. The researchers, however, would justify by claiming the "point reasoners" were merely giving *rote* "set" responses. In the same vein, the analyses for the data comparison probes (Figures 2.21 and 2.22) became quite problematic.

Figure 2.21 Same mean, different spread (SMDS) probe (Buffler et al., 2001, p.1147)[15]

Two groups of students compare their results for a distance measurement:

Group **A**: 444 432 424 440 435 Average = 435 mm
Group **B**: 441 460 410 424 440 Average = 435 mm

A: Our results are better. They are all between 424mm and 444 mm. Yours are spread between 410mm and 460 mm.
B: Our results are just as good as yours. Our average is the same as yours. We both got 435mm for the distance.
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. Do not use the word "results" in your explanation.

Figure 2.22 Different mean, same spread (DMSS) probe (Buffler et al., 2001, p.1147)

Two groups of students compare results for five releases of the ball at h = 400 mm:

Group **A**: 441 426 432 422 444 Average = 433 mm
Group **B**: 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. Do not use the word "results" in your explanation.

One example of a response to the SMDS probe in Figure 2.21 was as follows:

> They both got the same average and that's all that matters. It's not relevant whether the results are spread far or not.

The response below was one given to the DMSS probe in Figure 2.22:

> The two averages are so close that it is possible to say they agree with each other.

Both responses indicated the students were using only the mean value (which characterised "set thinking") to compare data sets but had ignored the spread. The students could not be categorised by the existing framework as their responses placed them in both paradigms. The researchers eventually decided the students were modelling an "*imposed* set reasoning" and were actually giving a rote response. Yet another problem surfaced when students who were initially identified as having an "*internalised* set reasoning" because they considered spread of data in the SMDS probe failed to do likewise in the DMSS probe. These students were then described as "*inconsistent* set reasoners" as

[15] The illustrations have been removed in these probes to save space. Only the text remained.

opposed to "*consistent* set reasoners" who would recognise spread in all data comparison probes. The need for constant refinements implied the categorising of students' thinking might not be as simple and straightforward as described by the "point/set" reasoning paradigms. The study by Lubben et al. (2001) underscored this fact when they proposed most students could be both "point/set reasoners" and which reasoning was to dominate might be dependent on factors like:

- "The purpose and complexity of task": for instance, students might use a "set paradigm" for interpreting measurements in the kitchen, but a "point paradigm" for those in the pharmacy/laboratory (Lubben et al., 2004). Apparently, the notion of greater accuracy required in the pharmacy or laboratory compared to the kitchen might spur students to suggest an "exact" value;
- "The context of the investigation": for instance, "point reasoners" might use repeated time readings (a "dynamic" measurement) to calculate the mean value, but for distance readings (a "static" measurement), they tended to look for a recurring value (Lubben et al., 2001);
- "Interference of prior knowledge": for instance, students might see calculating a mean as being a routine requirement of an experiment, and therefore, generated many readings in order to get one (Buffler et al., 2001).

The qualitative approach towards analysing the response data in these studies involved condensing the data into categories or themes based on valid inferences and interpretations. Looking at the way the responses were interpreted, several responses could have been evaluated too "harshly" and categorised as "point reasoning". First, when students' responses were

interpreted, there should have been more consideration given to their weak language abilities (which the researchers had reiterated a few times) that resulted in them poorly articulating their understandings of uncertainty. To illustrate, let us take a look at a response to the RT probe in Figure 2.20 (shown earlier):

> By releasing the ball more than twice from h=400 we can be more certain of our answer. If we release our ball maybe five times we can limit the chances of doing mistakes when using the stopwatch (Campbell et al., 2005, p.21).

Allie et al. claimed the response indicated repeating was a practice process towards a "perfect*"* measurement (i.e. a single value), and therefore, the respondent should be a "point reasoner". It is well-established the word "mistakes" has often been wrongly used to connote "experimental errors" (Taylor, 1997; Leach, 2002). Given the benefit of doubt, the student could have just meant the overall errors in using a stopwatch could be reduced by conducting preliminary trials, which is an acceptable procedure. The same "harsh" criteria were applied to another response shown below:

> You need to roll the ball a few more times because the first one or two measurements are usually rough estimates. You need to take more time measurements and then only can you take an accurate measurement (Campbell et al., 2005, p.32).

Lubben et al. (2001) claimed the respondent was a "point reasoner". As described in Section 2.2.6, we can reduce variation by being more careful and consistent in the way we handle measurements, and the suggested way of achieving this is through practice or by carrying out a few trial runs. Similar arguments can also be drawn against several descriptors of coded response derived for analyses of three probes (RT, RD, and RDA); two such descriptor-statements meant for "point reasoners" are given below:

- Practice will produce a more accurate or better measurement;
- You have to repeat until the readings are close together

(Campbell et al., 2005; p.105)

Another argument that must be considered in the categorisation of students' responses was whether the students might have interpreted the close-ended choice items differently (Ryder & Leach, 2000). Such a problem might happen because of the students' laboratory experience coupled with their poor grasp of the language used in the instrument. Allie et al. singled out responses as "point reasoning" whenever they appeared to indicate the pursuit of true values, for instance, the response below to the RT probe:

> The more measurements you take the more you know how accurate you are. One or two measurements don't tell you enough about the real time taken (Campbell et al., 2005, p.22).

The student's reason for taking more measurements was correct; the more measurements you take, the better the estimate of time would be (based on the concept of standard error). The phrase "real time" might not mean true value; instead, the student could be referring to the most accurate time that could be obtained. The strict interpretation by the researchers could also be seen in the analysis of response for other probes like the RDA probe shown in Figure 2.23.

Figure 2.23 Repeating Distance Again (RDA) probe (Campbell et al., 2005, p.93)

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 are not enough. We must release the ball several more times.

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

To this probe, one student responded:

> If the measurements are taken several times, it will be evident if the measurements correspond. It will be of great advantage finally to get the same measurement for several attempts (Allie et al., 1998, p. 452).

In a separate study, another responded:

> Since the ball was released again and two different results were obtained it is important to release it several more times until the equal or the same results can be obtained. This will be the exact distance required in mm (Lubben et al., 2001, p.317).

The researchers claimed that both responses indicated the repeats led to a recurring value which students perceived as the true value. But the responses could also be interpreted in terms of the students attempting to achieve a precise set of results. The precise readings might appear the "same" as a result of the limitations of the metre ruler (in terms of its resolution) in measuring the distance or the ball had landed on the "same" spot but with very small variations such that the points of contact on the floor could not be distinguished. To the students, the occurrence of recurring values indicated the measurements were highly reliable, and good enough to be used as evidence.

There were also problems in the coding descriptors. To illustrate, we shall look at the SMDS probe shown in Figure 2.21 again, and then relate to the descriptors used for coding the responses (Table 2.6).

Table 2.6 "Set" reasoning descriptors to SMDS probe (Campbell et al., 2005, p.111)

Option	Descriptor
C: I think that the results of group **B** are better than the results of group **A** because…	**B**'s results are closer together; they don't vary as much
	B's average is more accurate/reliable
	B's spread is smaller, so the average is more accurate

The basic purpose of the data comparison probes was to establish how the "quality of a data ensemble is characterised" (p.25). The researchers claimed that since the mean values for both data sets were the same in the SMDS probe, the spread should be the only factor to recognise the "the quality of a series of measurements" (Campbell, 2005, p. 38). If we examined the probe in Figure 2.21, group **B**'s results had a wider spread compared to **A**'s, and therefore, should not be better in terms of quality, and yet all three descriptors for option **C** in Table 2.6 classified as "set reasoning" referred to Group **B**'s data. The only way to correspond to such descriptors was for the

respondents to exclude the two values at both ends of group **B'**s data, which would be unacceptable given the values contributed to the spread of the data.

Data spread being an essential feature of "set reasoning" was further emphasised in other data comparison probes such as shown in Figure 2.24 below (and Figure 2.22 shown earlier).

Figure 2.24 Different Mean Overlapping Spread (DMOS) probe (Campbell et al., 2005, p.98)

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

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.

In these probes, the researchers claimed "the spread as an indicator of the uncertainty (standard deviation of the mean) of a set of measurements needed to be conceptualised and applied when deciding whether or not the intervals defined by the two series of data overlapped or not" (Campbell, 2005, p. 52). There are several issues with the design and analysis of the probes.

First, some responses to these probes that were eventually categorised as "set" were those that referred to the concept of range instead of variation. This was evident from a student's DMOS response given below cited as a "set response":

> Checking the spreading of Group **A** (424 to 444) and of Group **B** (438 to 462) you can tell that these two agree within the experimental error (Campbell et al., 2005, p.53).

However, the researchers did acknowledge later their inaccurate categorisation:

> The overall pattern of the responses suggests that the students were not able to differentiate clearly between the overall spread of the data ensemble and the differences between the individual data points within the ensemble. (Campbell, et al., 2005, p.26)

Second, the calculated SD and SE for both probes were high (for e.g., in DMSS, they were 7.07 and 3.16 for Group **A**; 9.43 and 4.21 for Group **B** respectively), which meant the degree of variation in the data sets of both probes was large, and rightly, there were insufficient data for the respondents to agree with any of the given options. Perhaps, an Option **C** could have been given for "we cannot tell" to cater to those who sensed the large variation in the data.

Finally, both DMOS and DMSS probes were about the "agreement" between two sets of DV data taken using a single value of IV, the height of 400mm where the ball was released from. This context seemed to be rather limited, and perhaps too contrived. The real and more important issue was to find out how much change could there be in the DV values with respect to uncertainties arising from changing the IV (i.e., with the same IV value, one would expect the true value of the DV to be the same, but with different values of the IV, one would not know and could only go with the data obtained).

Although the next two studies were conducted during laboratory investigations, they did not fully exploit the context to explore the understanding of uncertainties in DV repeats responding to a changing IV. This argument shall be expanded further later.

The first of the two studies were conducted by Séré et al. (1993). They studied the "processes of thinking when the students have to articulate the mathematical notions at their disposal with the practical and theoretical problems of measurement" (p.428). The research revolved around finding the focal length of a lens and the calculation of uncertainties in the experimental data. The key assumption was that the students' responses and comments in the laboratory report would reveal their understandings of uncertainty.

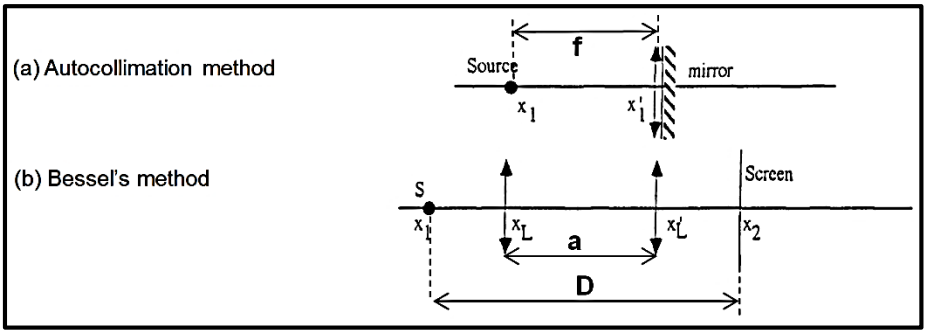
However, this was not to be: “If we consider initiative in making personal commentaries as an indicator, probably less than half of the students were in this category” (p.435). The description of the laboratory work in optics is encapsulated in Figure 2.25.

Figure 2.25 Laboratory work in Optics (modified from Séré et al., 1993, p.429)

Theme:	To measure the focal length of a converging lens by two methods: auto collimation and Bessel's method. The aim is not to elaborate a model of the experiment, by searching to see, for instance if there is one point where the light rays converge. The existence of the focus (point where light rays parallel to the optical axis converge) is postulated. The students have to find its position, though it is not a point.
Apparatus:	a light source, a converging lens, a white screen, a small mirror. All these components are arranged along an optical bench. Their positions are located on a metre rule graduated in mm. The coincidence between a component and a graduation on the metre rule is obtained with a plumb line.
Activity:	The students had a booklet at their disposal, designed to collect the results of their measurement, the answers to questions and the commentaries that they were asked for. It was also designed to guide the students' activities through the teaching sequence.

References were made to the autocollimation and Bessel’s method of finding the focal length(**f**) of a lens. Both are illustrated by the ray diagrams in Figure 2.26 and the methods of finding the focal length are described in Annex 2.1.

Figure 2.26 Methods of finding focal length(f) (Séré et al., 1993, p.429)



Sere et al.’s questions in the initial three pages dealt mainly with the concepts of accuracy and errors (see Figure 2.27). Using the autocollimation method, they required students to report a single measurement of the focal length and its uncertainty value. To get the latter, the students were expected to estimate the errors, but first, they needed to be able to identify different

sources of errors from the “measuring instrument, the experimenter, [and] the phenomenon itself” (p.431).

Figure 2.27 Laboratory questions (modified from Séré et al., 1993, p.430)

Page 1: Carry out one measurement of the focal length **f** of the lens by autocollimation. Give this value with an interval.

Page 2: Carry out a second measurement of **f**. Give it with an interval. If these two measurements are not completely identical, explain and comment on this fact.

Page 3: Do you think you are able to judge if one measurement is “good” (usable) and the other one less good? After these two measurements, are you able or are you not able to give a result? If yes, which result?

Page 4: Say how you can manage everything to get ten “usable” measurements. Carry out ten measurements.

Page 5: Can you give a value of **f**, with an interval, and the probability that **f** is within? Compare this result with the result given in page **1**, possibly with a graph. Any commentary is welcome.

Page 6: Note the ten values obtained by all of a group of students (ten values obtained from a hundred measurements). Following this sampling how is the manufacturer able to characterise the batch from which the lens come?

Page 7: In earlier laboratory work numerous values (**N**) of the same quantity (index of glass) had been obtained. Have you an idea of the reason why we did not make an average of the **N** values obtained?

Page 8: Carry out a measurement by Bessel's method. Give an interval for this value, after having calculated it by differential calculus, showing the contribution of the uncertainties on **a** and on **D** to the uncertainty on **f**.

Pages 9 and 10: The students are asked to make the measurements six times, to calculate the averages and the standard deviations of these quantities.
In which interval does **f** have 50% chance of being?
In which interval does **f** have 95% chance of being?
In which interval does **f** have 99% chance of being?

From their commentaries, the students seemed to be able to identify several errors in the focal length measurement:

> The lack of sharpness of the image, and the difficulty of the eye in evaluating it; the irregularities in the level of the supports; the thickness of the supports; the gap between the supports and the marks of their position; the lens being off the vertical; and the angle of the mirror to the optical axis (Séré et al., 1993, p.431).

However, only one group of students included *all* sources of errors in their calculation of uncertainty; the remaining nine chose to include only one that could come from one of the following:

- calibration error in the metre ruler used to measure distance;
- errors arising from the method of measurement, for instance, in finding the start and the end points due to the thickness of the lens and the lens support (see Section 2.2.4);

- errors due to the focal depth; the image formed was sharp over a *range* of distances ("There is a certain margin of distance in which the image is sharp. We try to find the best value around f 9.3<f<9.6", p.432) (See Section 2.2.5).

The students did not spontaneously repeat the focal length measurement despite claiming it was affected by errors. Most students only did their second measurement after being prompted by the laboratory instructions[16], which resulted in a slightly different measurement. Contrary to the researchers' expectation, the students did not attribute "chance [random] errors" to account for the difference between the first and second measurements; all they saw were systematic errors which they learned in their theory lessons.

Additionally, several students claimed they repeated a second time because they intended to check their "doubts about the first measurement" (p.432). These students put a lot of trust in their first measurement, and when they took the subsequent measurement, it was meant to evaluate the previous one and to replace it if necessary.

The researchers observed "only two [out of ten] groups suggested it would be relevant to carry out more than two measurements" (p.433). One could understand why the students did not intend to repeat their measurements. The distance measurements were taken using a metre ruler, a highly reliable instrument, and should not be much affected by uncertainties. As seen in the South African studies, students would probably expect the measurement of distance, a static measurement, to be highly repeatable. Even if there were variations in the repeats, they were expected to be small since all the measurements were taken with an identical set-up and no changes were made

[16] "They were expected to turn a page only when the preceding one had been completed" (p.433).

to any variable. Besides, if they were careful enough in their first attempt, any repetition would merely be an “imitation”, the idea of dismantling the set-up and re-assembling again to take another measurement appeared to be contrived and “profitless”.

After following instructions and obtaining several repeats, many students did not know how to handle the repeated readings; some proposed giving readings that appeared frequently a higher weighting. This led Sere et al. to claim that although the students held the notion “the more you make the measurements, the better the result is” (p.437), they did not clearly understand what “better” really meant. The students’ understanding of the purpose of repeats seemed to fit Perkins’s (2006, p.9) description of “ritual knowledge” as knowledge that transpired only “routine and rather meaningless” actions.

Looking again at the question on page three, it states: “After these two measurements, are you able or are you not able to give a result. If yes, which result?” (p.430); a question like this might be misleading as it implied two measurements could be sufficient. Besides, it seemed to contradict the idea of asking the students to find ten measurements later. Most responded by suggesting both data should be used to calculate a mean value, but some wanted to depict the values as a range or to pick the mid-range value; all these reflected the lack of understanding for finding a representative value for focal length. On the remaining pages, the researchers were mostly interested to find out whether the students could apply their statistical concepts like finding the confidence intervals from a batch of several focal lengths, and their statistical reasoning. Nonetheless, the students seemed to have problems applying their concepts:

> She was not able to…calculate the mean value and the standard deviation. Several had forgotten the basic definition of the standard deviation…The least important mistake was to write N instead of N-1 in the formula giving the

> confidence interval. The worst mistake was to write N-1 at the numerator, showing that the student did not understand that the formula expresses that increasing the number of measurements improves the precision. (Sere et al., 1993, p.435)

Some students might have developed misconceptions in their statistical ideas with one reponse describing SD as: "[allow*ing*] us to reduce as much as possible the random errors…"(p.435).

Sere et al. also investigated the students' conceptual understanding of key concepts like accuracy, precision, errors, and mean values in a final test using electricity as the context. There were not many details given in the report about the final test including its specific questions. Nevertheless, the few details revealed some ideas about its intent and findings.

The researchers found less than half their undergraduates could link SD to the concept of precision, and only about half acknowledged the contribution of *both* random and systematic errors towards uncertainty in all measurements, the rest would only impute only one type of error. Additionally, when asked to account for the differences in accuracy or precision of results, many students were unable to do so; instead, they gave frivolous comments such as "bad apparatus", "a lot of errors", "differing measuring procedures", and "the instruments are different for each experimenter, so it is impossible to find the same results" (p.436). Another problematic concept was the mean value. Some students viewed the mean value as a theoretical value, and expressed the idea of "the average of the measurements of the diameter of the metallic wire under study is not the real wire but the diameter of a hypothetical wire which would be perfectly cylindrical" (p.434).

The test also sought to examine students' idea of the quality of measurements by asking them to compare a set of experimental data (with

uncertainty values) to a given value. The researchers commented the students were often using unclear qualitative terms:

> …the students often used the terms 'good/bad measurement' and even more often 'better measurement'. They established a sort of hierarchy between the measurements, instead of taking into account the whole set (Sere et al., 1993, p.436).

Perhaps, the students had taken the cue for making such "unclear" comments from their coursework, which is evident on page three of the laboratory manual (Figure 2.27) where the term "good" was used to compare the accuracy of focal length measurements.

Looking at the use of the laboratory in the study to elicit students' understanding of uncertainty revealed several problems. The use of laboratory questions as the main source of evidence depended on the students' active response. The researchers had given brief instructions and little scaffolding (see Figure 2.27) understandably because they were dealing with undergraduates, and presumably, they did not want to influence the participants' response by the questions they asked. However, "few students showed…initiative...to make personal comments on the results, [and] giving some feedback on the intermediate results"(p.437).

A critical observation was that many students were unable to demonstrate their conceptual understanding in the laboratory investigation because of their "poor understanding of the *procedures*" (p.437). This implies a serious disadvantage of using laboratory investigation as a context for researching understanding as it hinges on the participants being able to effectively apply their manual skills and knowledge of experimental procedures to perform the investigations.

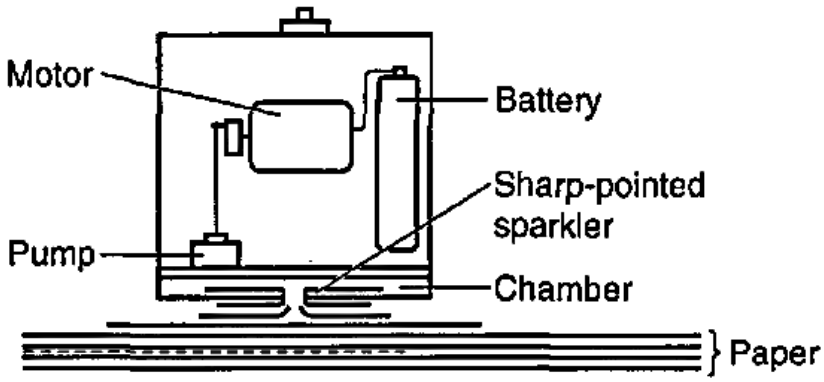
The next study by Coelho and Séré (1998) looked at students' *constructed* meanings about measurements in a scientific task via two sets of

45-minute clinical interviews with groups or individuals as they were performing their investigations. According to the researchers, the interviews allowed them to "produce the analysis not only of the 'saying' but also of the 'doing' " (p.81).

The first set involved ten students responding mainly to questions pertaining to data collection, thus, the interviews were known as collection interviews (CI). The second involved interviewing the remaining eleven students mostly on data processing and interpretation, so the interviews were labelled as processing interviews (PI). The decision for separate interviews was taken to avoid long interview sessions[17]. The use of interviews may be good at clarifying underlying reasons for students' actions but the researchers have to be cautious that their "very act of asking about reasons for actions might be taken as a sign that the group was doing something wrong" (Millar et. al., 1994, p.212). The corpus of research data also came from the students' completed laboratory reports. The researchers adopted a more interpretivist approach to making sense of their evidence; they did not use inferential statistics as they perceived the sample size as being too small to be of any significance.

The investigative task involved measuring the uniform velocity of an "air puck" that moved along a horizontal table on an air cushion after it was given a gentle push (Figure 2.28).

Figure 2.28 The "air puck" set-up (Coelho & Séré, 1998, p. 95)



[17] The students, however, were allowed to complete the whole investigation, and not dictated by the focus of the interviews either being on collection or interpreting.

The recording of the puck positions was carried out when the air puck reached uniform velocity, which in turn was indicated by recordings of a sharp pointed sparkler that produced dark burnt dots at regular intervals (every 20 milliseconds). Examples of recordings (with some students' annotations) are shown in Figures 2.29 and 2.30.

Figure 2.29 A "CI" recording with students' annotations (Coelho & Séré, 1998, p.83)

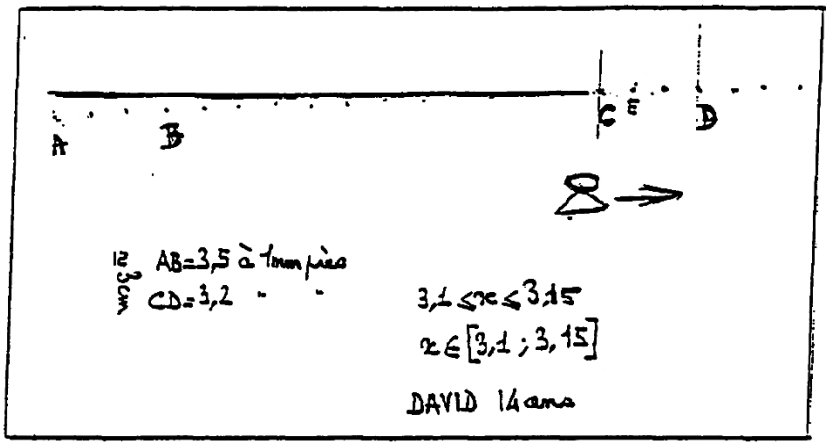
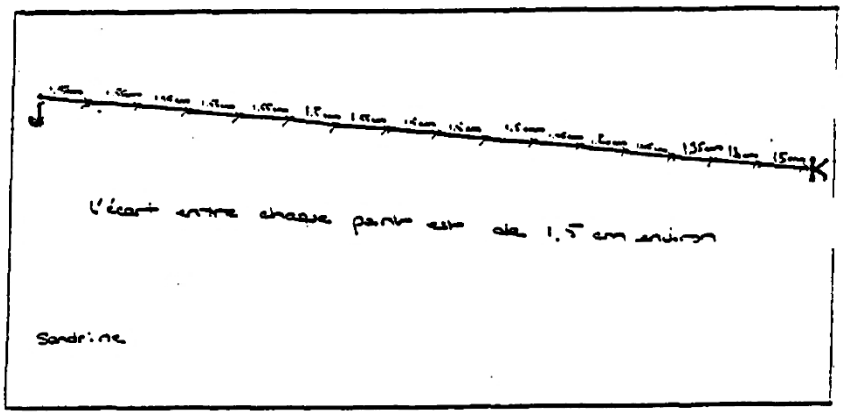


Figure 2.30 A "PI" recording with students' annotations (Coelho & Séré, 1998, p.83)



There were no "fixed" questions for the interviews rather the researchers' questions depended on the students' actions in their investigation. Looking at the objectives and sample questions in Table 2.7, one can see that Coelho and Séré's questions were basically focused on uncertainties arising from finding the distance between two points (single measurements) or finding distances between different points (repeated measurements). Our following discussions will therefore revolve around the same emphases.

Table 2.7 Summary of interview objectives and questions (modified from Coelho & Séré, 1998)

Collection Interviews (CI)
Interview objectives:
• Making students face a situation which gives rise to measurement practice. • Seeing how the student responds regarding uncertainty, on account of the fact that dots on the recording have a finite size. • Seeing how the student defines the measurand and the measurement procedures. • Seeing how the students' interpolation process occurs, and how he/she responds in face of the reading uncertainty. • Seeing student's response in the presence of the different results: checking whether there is the idea of 'true value'. • Surveying the ways the data are communicated. • Checking whether students consider the measurement uncertainty, and see which the criteria were adopted to justify the way the result was expressed, as well as the number of digits maintained.
Sample interview questions:
• What can you say about the average velocity of the puck between **A** and **B**, and between **C** and **D**? (See Figure 2.29). • Could you say how you have made the measurement? Where did you place the ruler? • Where did you exactly place the zero of the ruler? Which references did you choose for the measurement? What does 'between 3.1 and 3.2' mean? How much? Why do you round off figures? Can you repeat the measurement for the distance **CD**? • What do you think of this? Why have different results been obtained? Is there any result which is better than the other? If you had to express the measurement result, how would you do it? Why do you use the average? Why do you keep four digits after the point?
Processing Interviews (PI)
Interview objectives:
• Seeing how students use the recording spontaneously, which role the measurements play in relation to other aspects, such as students' visual perception or their ideas concerning motion. • Seeing which criteria students use in order to decide whether the measurement results are compatible. • Seeing how students explore the measurement results: whether they need the help of tables, graphs, whether they search for a variation law, whether they consider the uncertainty of the measurement. • Seeing students' previous ideas and their knowledge of physics interfere with the analysis of the results.
Sample interview questions:
• Does the puck velocity between **J** and **K** vary? Justify your answer. • Describe all the steps you have followed in order to answer the above question. • If you had to tell someone of the result of the measurement, how would you express it? • First 1.5 cm, then 1.55 cm, and then 1.4 cm and 1.5 cm. Why do you think there is such a difference? Has the velocity decreased? • First you said you were going to make a scale, then you said you were going to measure again in order to find the same value, and then be able to decide and finally you said 'well, we cannot produce measurements which are all equal. • Does the variation observed derive from 'measurement errors', as you mentioned, or is it due to velocity variation? • From the result of these measurements, what allows you to say that velocity has decreased?

The uncertainties related to a single distance measurement were largely surfaced during the CI phase where the questions dealt mostly with the choice of measuring instrument (a ruler graduated in millimetres) and the method of taking measurements. With respect to the ruler, the errors raised were mainly concerned with its resolution of scale (thus, “reading errors”) as the

students experienced difficulty in reading distance measurements marked by burnt marks that appeared between two divisions of the ruler. They claimed the interpolation of measurements was quite arbitrary and inaccurate, and therefore, suggested replacing the metre rule with a more accurate measuring instrument.

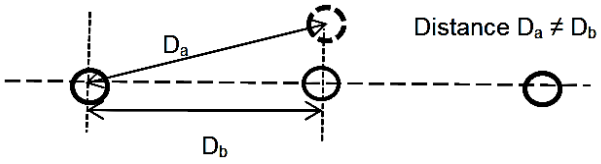
The researchers also noted a certain proportion of their students often used terms such as “exact”, “correct” or “fair” to express the idea of a true value. Coelho and Séré claimed their students believed in the possibility of getting a true value either by using better quality instruments (that the students claimed were unavailable in their school) or by taking precautions in the measurement process such as “preventing hands shivering” (p.87). The researchers concluded the students could have developed a “myth of the physicist and/or the perfect measuring instrument” (p.87). The idea of a “perfect” instrument was likewise observed by Leach et al. (1998), and more recently by Munier et al. (2012).

On the method of taking measurements, the students were generally concerned with the different sources of random errors such as:

- Finding the start and end points for distance measurements. The sample CI questions in Table 2.7 implied several “dilemmas” in this area. First, the dots marked on the recordings could be of different dimensions and shades, which led students to think of different beginning and end points. Second, the sparkler occasionally “generated double sparks and burnt the paper in an irregular way, giving traces rather than dots” (p.82), which led to various proposals for taking distance measurements, for example, between the extreme left or the extreme right of the traces.

- Errors inherent in the measurement procedure, for instance, several students cited the thickness of the pen markings to indicate the points on the recordings, and that the dots were not linear. On the latter, the misalignment of the dots provoked a particular CI question: “Must the measurement be made on a trajectory previously defined as rectilinear and drawn taking a point from the dots as a reference?” The misalignment was a source of error as the measurements between two dots might be different (see Figure 2.31; as indicated, the distance Da of a misaligned dot would not be equal to distance Db given by a dot formed in a straight line).

Figure 2.31 Errors in measuring distances between misaligned dots



From Table 2.7, the uncertainties related to repeated measurements were mainly surfaced in the PI. Herein, the students were questioned on their recordings; in particular, whether it showed constant, increasing or decreasing velocity (the velocity could be determined *directly* by comparing the distance between two successive dots that were recorded at constant time intervals). Technically, a constant velocity would be shown by equal distance measurements between successive dots, but this was not seen in the recordings (see Figure 2.30). The interviewer then took the opportunity to tease out the students’ understanding of uncertainty by posing several questions with reference to their recording (Figure 2.30):

- Does the puck velocity between **J** and **K** vary? Justify your answer.
- Describe all the steps you have followed in order to answer the above question.
- If you had to tell someone of the result of the measurement, how would you express it? (Coelho & Séré,1998, p.82)

Several other sources of uncertainty were raised when interpreting the measurements. For instance, when students were asked to determine the overall velocity, the researchers found the students' methods varied in the way they measured distances. Table 2.8 shows students taking measurements in different ways (with reference to Figure 2.30).

Table 2.8 Variation in methods of finding overall velocity (Coelho & Séré, 1998, p.91)

1. Measurement of the distances between two non-consecutive points, and determination of the instantaneous velocity at an intermediate point
2. Measurement of several distances at the beginning, in the middle, and at the end of the recording.
3. Measurement of several non-successive distances between two successive dots at the beginning, in the middle, and at the end of the recording.
4. Measurement of the first and the last distance.
5. Measurement of the distances with the help of a pair of compasses, bringing them on to a graduated ruler.
6. Measurement of all the distances between two successive dots.
7. Drawing of straight line segments; two measurements at the beginning and two at the end.
8. Ruler reading, providing interesting results, i.e. equal distances.
9. Measurement of the total distance d_{JK} and division of the result by the number n of distances (d_{JK}/n).

Incidentally, one group found all their distances decreasing, and this led them to interpret the overall velocity had decreased. The response below from one student illustrates this point:

> Intervals change anyway. The interval [d_l] was 1.5cm and 1.4cm, and here [d_i] it is l.4cm and 1.3cm. Really, there is an important decrease in velocity since the intervals [d_i] have changed...There is always a 1mm interval range, but it is not between 1.4cm and 1.5cm anymore; it is between l.3cm and 1.4cm. (Coelho & Séré, 1998, p.92)

The students then justified their conclusion by citing the effects of table friction and air resistance. This led the researchers to suggest the misperception was reinforced by the "confusion between measurement uncertainties and discrepancies in the expected results as predicted by theory" (p.93). The group

did not realise the decrease in the selected distances was actually a random event. There was, however, nothing in the report to suggest that a follow–up action (perhaps, at the end of the PI) with the group was done. From an education perspective, the interviewers could have asked the students to account for the errors and what could be done to improve the reliability of their measurements. The students might then be led to identify the sources of errors such as uncontrolled variables (for e.g., the table friction and air resistance that the students cited earlier) and to improve reliability by suggesting more repeats (i.e., longer recordings) to study the pattern of the results. Alternatively, the students could be prompted to check the unselected distances in their existing recording.

On the whole, the study by Coelho and Séré showed the students had intuitive ideas about uncertainty in measurements. However, the researchers did not go far enough to investigate this especially with respect to the effects of uncontrolled variables. Besides, the understanding of uncertainty in measurements could have been more meaningfully investigated if the focus of investigation was on finding the relationship between a continuous DV and IV (for e.g., the effect of increasing the load on the puck on the velocity).

On reflecting the use of laboratory investigations to explore students' understanding of uncertainty, I realised it might take a longer duration and would be labour-intensive especially if close observations were required. Besides, two factors might pose further challenges for my own research: first, the limited access to PSTs, and second, the unavailability of laboratories to conduct the research. Even if these problems could be overcome, it would be logistically cumbersome to set up investigations and to schedule the PSTs (as well as make provisions for those who missed), Besides, there would be a

whole host of validity and reliability issues related to research data collection during a laboratory investigation (for e.g. prompting students during an investigation). Additionally, having to deal with groups as seen in the study by Coelho and Séré might also invalidate the research findings if the PSTs were to collude in response to the questions. Finally, as Kanari and Millar (2004) had noted, even in tightly defined investigations, students' actions as well as the quantity and nature of the data collected might vary so widely that it could be difficult to identify regular patterns in their actions, and to generalise their understandings.

Despite all these shortcomings, laboratory-based research might be the only "possible method to explore adequately some aspects of procedure (such as the judgement of significance of small differences between measurements)" (Lubben & Millar, 1996, p.955). Besides, uncertainty could not have been fully dealt with in isolation (by using a questionnaire or an interview) as one might have to consider many different aspects of an investigation related to planning, performing, interpreting and evaluating measurements.

Based on the review, we could see to a certain extent an attempt at triangulation in the study by Séré et al. (1993) as the researchers deployed different methods (laboratory investigation and written test) and different topics (optics and electricity), as well as in the South African studies, where refinements were carried out on the questionnaires administered to different groups of students over different years. However, much of the triangulation efforts were directed towards confirming initial observations rather than checking interpretations of students' responses.

As seen in the South African studies, the use of the written questionnaire enabled the researchers to reach many participants, and they

could revise their instruments several times quickly to reveal patterns and divergences in the students' understandings of uncertainty in measurements (which would have been a strong point if the intention was to check the interpretations of students' response). But there were disadvantages as well. It depended on the respondents being able to articulate their understandings, and if the respondents have poor grasp of the language of the questionnaire, as in the case of the South African studies, getting quality research evidence might be impeded.

Further, as seen in the study by Evangelinos et al., it might be difficult to interpret students' understandings through their written responses especially if they lacked details, vague and ambiguous. Moreover, the responses could not be further clarified unless some additional steps like a follow-up interview were arranged. In addition, the questions as well as the coding scheme could not be modified in the light of new evidence and interpretations, particularly if the instrument was to be deployed again, as this would seriously violate its validity.

To a certain extent, the use of interviews might reduce the problems inherent in the use of a questionnaire. Leach et al. (1998) claimed an interview study would be able to yield rich insights into the students' ideas of uncertainty in data. Understandably, the qualitative approach permits a researcher to explore and uncover the participants' ideas and actions in greater depth as it allows the researcher to seek information and clarification from the participants (and thereby, improves interpretation).

Finally, another important aspect drawn from the review was the prevalent use of probes across the literature, and in many large scale studies including the PACKS project (Millar et al., 1994; Lubben & Millar, 1996) and the Labwork in Science Education project by the European Commission (Leach et

al., 1998). The use of probes was premised on the assumption that the understandings required to carry out a practical task were substantially knowledge-based, and therefore could be articulated. However, this did not mean all procedural knowledge could be made explicit as the choices and decisions to combine different ideas into an overall strategy might be based on tacit understanding (Lubben & Millar, 1996). The case for supporting the use of probes was well put by Lubben and Millar when they wrote:

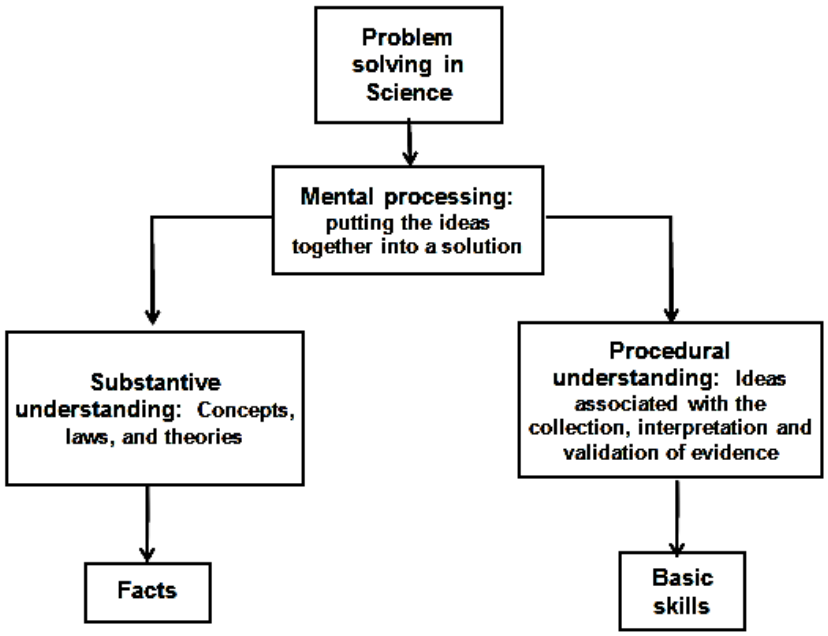
> …there is some value in exploring students' ideas about some specific aspects of scientific procedure in isolation. This might be construed as a model of performance in which competencies on discrete sub-tasks are seen as setting the upper limit of performance on whole tasks…What diagnostic probes do offer, however, is the possibility of getting relatively large samples of students to respond to identical stimuli in a way which is impossible to engineer in a practical investigation setting.(p.957)

A critical factor, according to the study by Millar et al. (1994), in the performance of scientific investigations including handling uncertainty was the "ideas which underpin [the] criteria for evaluating the quality of empirical data (understanding of evidence)" (p.207). The next section deals with this factor; specifically, it describes the theoretical framework that guided this research to identify "ideas" the PSTs used in understanding uncertainty in measurements.

2.5 Concepts of Evidence (CofEv) as a Theoretical Approach

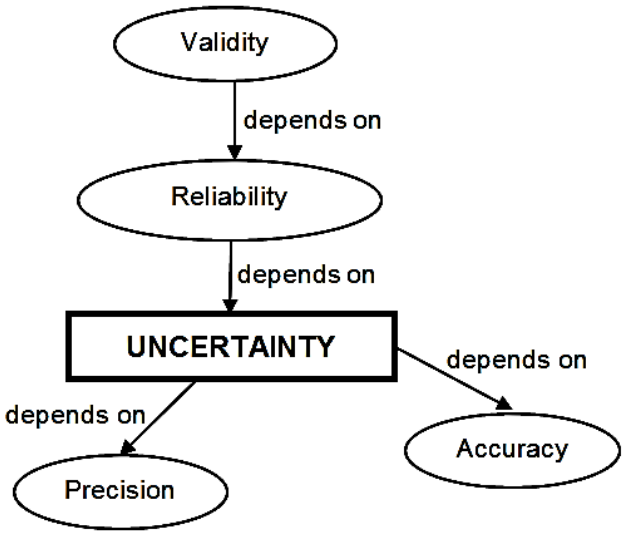
The CofEv represents a domain specific area of procedural understanding that functions alongside substantive understanding in tackling problems in science, and importantly, the CofEv underpin the key concepts of validity and reliability (Gott et al**.,** 2008) (see Figure 2.32). According to Gott and Duggan (2003), if measurements were unreliable then the whole investigation automatically became invalid because the measurements could not be trusted, and therefore, the interpretation of results and the conclusion could not be trusted as well. Thus, it all boils down to the measurements.

Figure 2.32 Role of procedural understanding in solving science problems (from Gott et al., 2008)



Gott et al. (2008) emphasised that CofEv included "ideas about the uncertainty of data" (p.13); and, "precision and accuracy are fundamental to all measurements and underpin reliability" (Gott and Duggan, 2003, p.120). In other words, both concepts that underpinned the understanding of uncertainty would allow us to decide whether a measurement was good enough to be accepted as evidence. Figure 2.33 illustrates how uncertainty in measurements can be related to CofEv that underpinned the overarching concepts of validity and reliability.

Figure 2.33 Linking uncertainty to reliability and validity (modified from Gott & Duggan, 2003)



A critical understanding implied in Figure 2.33 is that if we take the extreme case where the measurements are imprecise or inaccurate (i.e., the size of uncertainty is too large), then the reliability of the measurements will be called into question, and this eventually leads to the invalidation of the whole investigation.

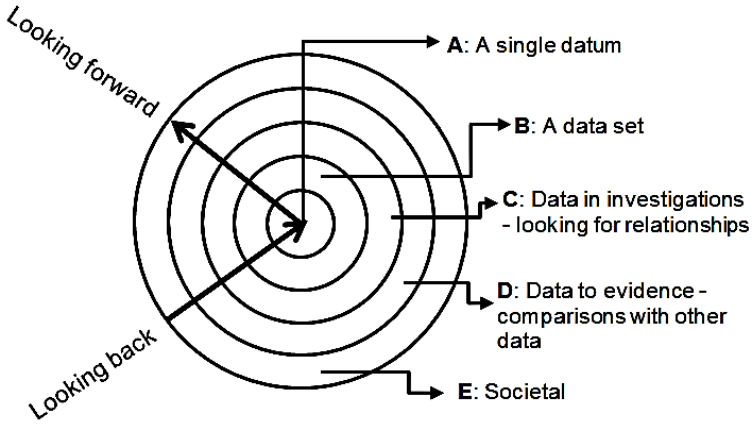
2.5.1 What are CofEv?

Essentially, the CofEv adopts a constructivistic view towards the use of ideas or concepts in handling/solving "problems" (including uncertainty) in an investigation. In a science investigation, CofEv serve as a "toolkit" of ideas that learners can use in their decision-making process at any stage of their investigation; their practical actions therefore will be executed with understanding rather than as a routinised procedure (Roberts & Gott, 2003). In other words, a learner who really understood evidence in terms of the concepts that underpinned validity and reliability would be able to apply their understanding in tackling issues, including those concerning uncertainty, in an investigation. Gott et al. (2008), however, cautioned that CofEv alone may be insufficient to complete an investigation, substantive knowledge as well as manipulative skills would also be necessary.

Gott and his co-workers viewed data as a primary concern in any student's evaluation and decision-making processes in an investigation; thus the notions of reliabilty and validity are built around collecting quality data starting from the act of sampling a single datum, through collecting a data set and presenting the data as evidence, to using the evidence in making and defending scientific claims. The bullseye diagram shown in Figure 2.34 represents the ways in which the CofEv are factored in an investigation (the

letters **A** to **E** do not connote that the bullseye diagram should be applied in a linear sequence, rather it is just a way of categorising the CofEv).

Figure 2.34 Bullseye diagram of the CofEv underpinning validity and reliability (Gott, et al., 2008)



Referring to the bullseye diagram, Gott et al.,(2008) pointed out that when an investigator is seeking a conclusion by making an evidence-based claim("looking forward") or when the investigator is checking on the evidence-based claims of others("looking back"), he or she needs to have the goal of producing valid and reliable data borne in mind at every layer starting from the very first. Essentially, the bullseye diagram should be viewed in such a way that each preceding layer is nested within the subsequent layer irrespective of whether the investigator begins his or her task from **A**("looking forward") or **E**("looking back"). In other words, the understandings of the ideas of evidence nested in all layers **A** to **E** would be drawn upon in various combinations when the learner is "looking forward" or "looking back". This means how the learner constructs and establishes the validity and reliability of evidence in an investigation is eventually dependent on the multiple interactions of the CofEv including those that are related to uncertainty in measurements.

Building from their research on procedural understanding, Gott et al. identified a list of 80 or so CofEv (see Annex 2.2) which they believed could serve as a knowledge base for learners to understand scientific evidence and

help them make decisions in justifying claims or in counteracting others. The authors claimed the list is dynamic and is constantly being refined and updated to meet evolving science education trends and needs.

An overview of the CofEv will show that the concepts are structured into six categories (see Table 2.9), which are then subdivided into 21 areas ranging from fundamental ideas in measurements and instruments, through ideas on collecting, analysing and interpreting a single datum or a set of data, to ideas related to validity and reliability as the principal criteria for evaluating the quality of evidence, and ideas that are related to relevant societal aspects such as credibility and practicality.

Table 2.9 Categories and areas of procedural understanding within the Concepts of Evidence framework (Gott & Duggan, 1995)

Categories/Relating to the bullseye diagram (Figure 2.34)	Areas of procedural understanding
Core ideas	1. Fundamental ideas (e.g. how students weigh opinion and data as evidence?) 2. Observation (e.g. how students see objects and events using their substantive understanding?) 3. Measurement (taking the presence of inherent variation into account)
Making a single measurement/ Layer **A**	4. Underlying relationships in an instrument which converts the variable being measured into another that is easily read 5. Calibration and error in the measuring instrument 6. Reliability and validity of a single measurement
Measuring a datum/ Layer **B**	7. The choice of an instrument for measuring a datum 8. Sampling a datum 9. Statistical treatment of measurements of a datum 10. Reliability and validity of a datum
Data in investigations - looking for relationships/ Layer **C**	11. Design of investigations: Variable structure 12. Design of investigations: Validity, 'fair tests' and controls 13. Design of investigations: Choosing values 14. Design of investigations: Accuracy and precision 15. Design of investigations: Tables 16. Reliability and validity of the design 17. Data presentation 18. Statistics for analysis of data 19. Patterns and relationships in data
Data to evidence - comparisons with other data/ Layer **D**	20. Reliability and validity of the data in the whole investigation
Societal issues/ Layer **E**	21. Relevant societal aspects

To illustrate the use of the CofEv, the CofEv associated with one of the 21 areas that is sampling a datum is shown in Table 2.10. As shown by Table 2.10, each of the CofEv is defined by how it is understood in the context of an investigation, and then exemplified so that each concept has greater clarity and a meaningful description.

Table 2.10 CofEv associated with sampling a datum (modified from Gotts et al., 2008)

CofEv	Understanding that…	Example
Sampling	One or more measurements comprise a sample of all the possible measurements that can be made.	A single measurement of bounce height of a rubber ball is a sample of the infinite number of such bounces that could be measured.
Size of sample	The greater the number of readings taken, the more likely they are to be representative of the population.	The more times the ball is bounced and its height measured, the more likely the sample represents all possible bounces of that ball.
Reducing bias in sample/ representative sampling	Readings must be taken using an appropriate sampling strategy such as random sampling, stratified or systematic sampling so that the sample becomes as representative as possible.	The ball that is selected can be any of the balls available; no preference is given to any particular brand of ball.
An anomalous datum	An unexpected datum could be indicative of inherent variation in the data or the consequence of a recognised uncontrolled variable	A very low rebound height from the ball may occur as a result of the differences in the material of the ball and is therefore part of the sample.

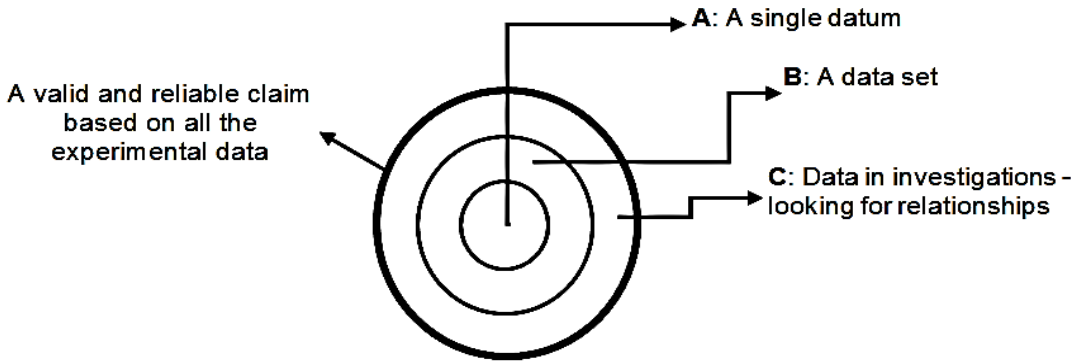
Note: The term “sampling” means any sub-set of a “population”. Since the examples are concerned with the investigation of the first rebound height of a ball, the term “population” refers to an infinite number of repeated readings taken for rebound height.

2.5.2 Scoping the CofEv to this Study

The bullseye diagram shown in Figure 2.34 describes the application of CofEv in an investigation that leads to a *public* claim, and such endeavour can be an outcome of a scientist or a university research team. According to Gott et al. (2008), for the majority of school-based investigations, the CofEv that are relevant are those associated with the innermost three layers of the bullseye diagram (Figure 2.34). At the primary level, investigations typically do not end up with making public claims; rather, most investigations are likely about finding a final conclusion to an investigative problem, for instance, to establish the

relationship that existed between two or more variables (see examples in Section 1.4)

Figure 2.35 Bullseye diagram of school-based investigation (Gott et al., 2008)



As seen in Figure 2.35, the most relevant layers are **A**, **B** and **C**. Let us look briefly at each layer individually[18].

Layer **A** has to do with the validity and reliability of making a *single measurement* of any variable that needs to be measured either quantitatively or qualitatively. The CofEv within the layer are concerned, for example, with the range and sensitivity of the measuring instrument, how the variable can be derived from other measurements (for instance, distance and time in the case of speed), and whether the chosen instrument and method do in fact measure the quantity we want in a valid way. Once the single datum has been obtained, the CofEv in the layer looks at its accuracy, and to see whether it can be estimated simply by considering the instrumental accuracies. We then have to decide if one measurement is enough to meet the purpose of taking the measurement in the first place; if not, repeats will then be necessary.

Layer **B** has to do with a data set of a variable, and the validity and reliability of the *repeated measurements* in the set. The CofEv within the layer are concerned about how to establish the degree of variation between

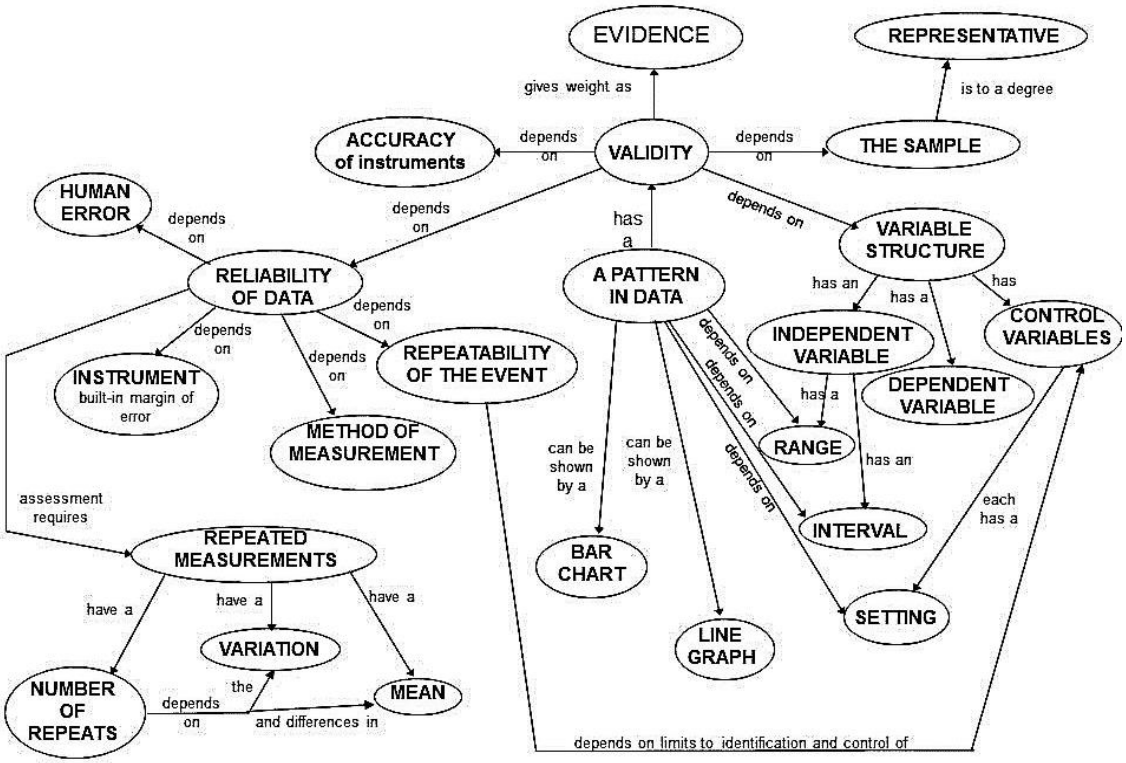
[18] Due to space constraint, a complete account cannot be given. For this, we can refer to Annex 2.2.

successive measurements, which in turn demand decisions on the number of repeats necessary to establish a certain confidence level suited for the purpose of the measurements. This can be done through logical reasoning, or if the data is sufficient, by using statistical methods such as the calculations of the mean value and standard error. The analyses may lead to further decisions on the number of readings to reduce the standard error.

Finally, the CofEv in Layer **C** are concerned with the overall design of the investigation, its validity and reliability as well as establishing the relationship between one or more of the variables by looking at the patterns in the data, that include the changes in the *repeated measurements of a DV in response to a changing IV*. In designing the investigation, the learner needs to consider the identification and effects of the CV, the range that is needed to establish any potential trend, the interval between the IV readings to "catch" the maximum and minimum points in a relationship, and what will be a sufficient number of readings to determine the pattern between the DV and IV. After the data have been analysed, we must get some sense of the range of errors and variation in all the readings in order to be confident about the validity and reliability of the observed trend line or curve. The concept of preliminary trials is also important as trials can give "a feel" for the range and interval of the IV (besides trying out the measuring instruments or the method of measurement); Johnson (2013) suggested conducting trials at the beginning and the end of the scale for the IV to see whether the scale needed to be expanded further in order to determine the full extent of the relationship between the variables. Such trials can also be extended to the number of DV repeats that correspond to the first and final values of the IV in order to get a sense of the degree of variation in relation to the magnitude of the change in DV for different values of the IV.

Describing the concepts according to layers **A**, **B** and **C** of the bullseye diagram (Figure 2.35) helps us to see the relevant CofEv associated with each layer and their contributions to understanding uncertainty. The CofEv nested within each layer are interacting with each other and with those in other layers in order to establish the overarching concepts of validity and reliability of evidence. The understandings of how these concepts operate individually and collectively would enable the student to check the quality of evidence as he or she "looks forward" in a science investigation. Figure 2.36 below from Johnson (2013) gives a good overview of the CofEv that are nested in the three layers of the bullseye diagram in Figure 2.35.

Figure 2.36 Overview of CofEv in primary science investigations (Johnson, 2013)



What the concept map in Figure 2.36 also shows is the multiple interactions of concepts or ideas to establish the validity and reliability of evidence. Since the overarching concepts include the concept of uncertainty in measurements, the same concepts that underpin validity and reliability that we

see in Johnson's concept map may also be required in handling uncertainty in measurements. Applying the preceding notion to the PSTs, to be "leaders of inquiry", and in particular, to be able to handle measurements and their uncertainties properly and to avoid those problems described in Section 1.5, one needs to have a good understanding of the CofEv including knowing how to apply and synthesise the different CofEv effectively.

2.5.3 Applying the CofEv

Following the discussions from Sections 2.5 to 2.5.2, the study proposed the use of the CofEv to represent the knowledge-base of ideas for PSTs to understand uncertainty in measurements (as described in Section 2.2: "The Science of Uncertainty in Measurements") but keeping in mind that often a combination of ideas might be necessary to handle the uncertainties in measurements taken during science investigations. Granted with this premise, the CofEv would therefore be used for:

- the construction of items for the research instruments in this study;
- the analysis of response to the items in the instruments.

However, considering the limited amount of time and space given for this thesis, it will not be in its interest if the study chooses to explore all procedural ideas expected of a PST to demonstrate understanding of uncertainty in the different measurements taken in a science investigation. At the risk of the study having too wide a focus and too little a time to achieve its goals, it chooses instead to focus on exploring a number of *basic* procedural ideas that are needed in understanding uncertainties in different measurements taken in a primary science investigation (See Table 2.11). Table 2.11 was derived partly from the review of literature in Section 2.4, and partly from the review of the primary science syllabi of England (Department of Education and

Skills [DES], 2013), Australia (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2013), and Singapore (MOE, 2008, 2009).

Table 2.11 Basic procedural ideas investigated in this research

Categories	Basic ideas
• Core ideas	The presence of inherent variation in all measurements; Concept of true value; Concepts of accuracy, precision, and experimental error
• A single measurement	Accuracy of instruments
• Repeated measurements of a single quantity	Purpose of repeats; Causes of variation; Anomalous result, Number of repeats to judge reliability; statistical treatment of repeated measurements using the use mean, range, standard deviation and standard error; Instrumental reliability
• Repeated measurements of a DV changing with an IV	Patterns and relationships in data; choosing values; data presentation in Tables

The list, however, is not intended to specify the syllabus requirements for primary students in these countries but simply to state several basic understandings required of a primary science teacher in handling uncertainty in measurements. Certain ideas like SD and SE were included as the study assumed teachers would need to have more than the specified content knowledge than their students (as proposed by Shulman, 1987). On the other hand, certain ideas including data presentation of graphs and charts were excluded, not because they were not critical, but due to the lack of time and space (in this thesis) to deal with them in depth, and the decision to leave them to a future study.

2.6 Summary of Chapter

Chapter 2 developed the methodological basis for eliciting the PSTs' understanding of uncertainty in measurements, and that is to strive for a "neutral ground" in the interpretation of the PSTs' response to the researcher's questions. The chapter began by covering the science of uncertainty in measurements which involved the conceptual understanding of accuracy, precision, and errors, as well as how they are interrelated. The relevant literature on uncertainty in measurement was reviewed to see how other researchers had sought to investigate the understanding in their subjects. A wide array of research strategies could be seen; the learning points gathered from the review would contribute towards developing the methodology for this current research. In addition, the findings could also be used to compare with those from the current research and to draw inferences. Finally, the chapter also presented its theoretical framework based on the Concepts of Evidence to achieve its research aims. The CofEv are viewed as a knowledge base of procedural ideas that are used to handle uncertainty in measurements, and can be used as a framework for designing the instruments and analysing the responses.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Chapter Overview

This chapter explains the methodology for this research. It will provide the rationale for its qualitative approach towards an accurate interpretation of the PSTs' understanding of uncertainty in measurements, and includes reasons for using different instruments for data collection. Further details about the participants will be provided along with the measures taken to safeguard their interests. It will also describe how the data are analysed and processed for evidence; these involved reducing the data into analysable units, coding, and categorisation. At different points in the chapter, issues on qualitative validity and reliability (Creswell, 2009) will also be addressed.

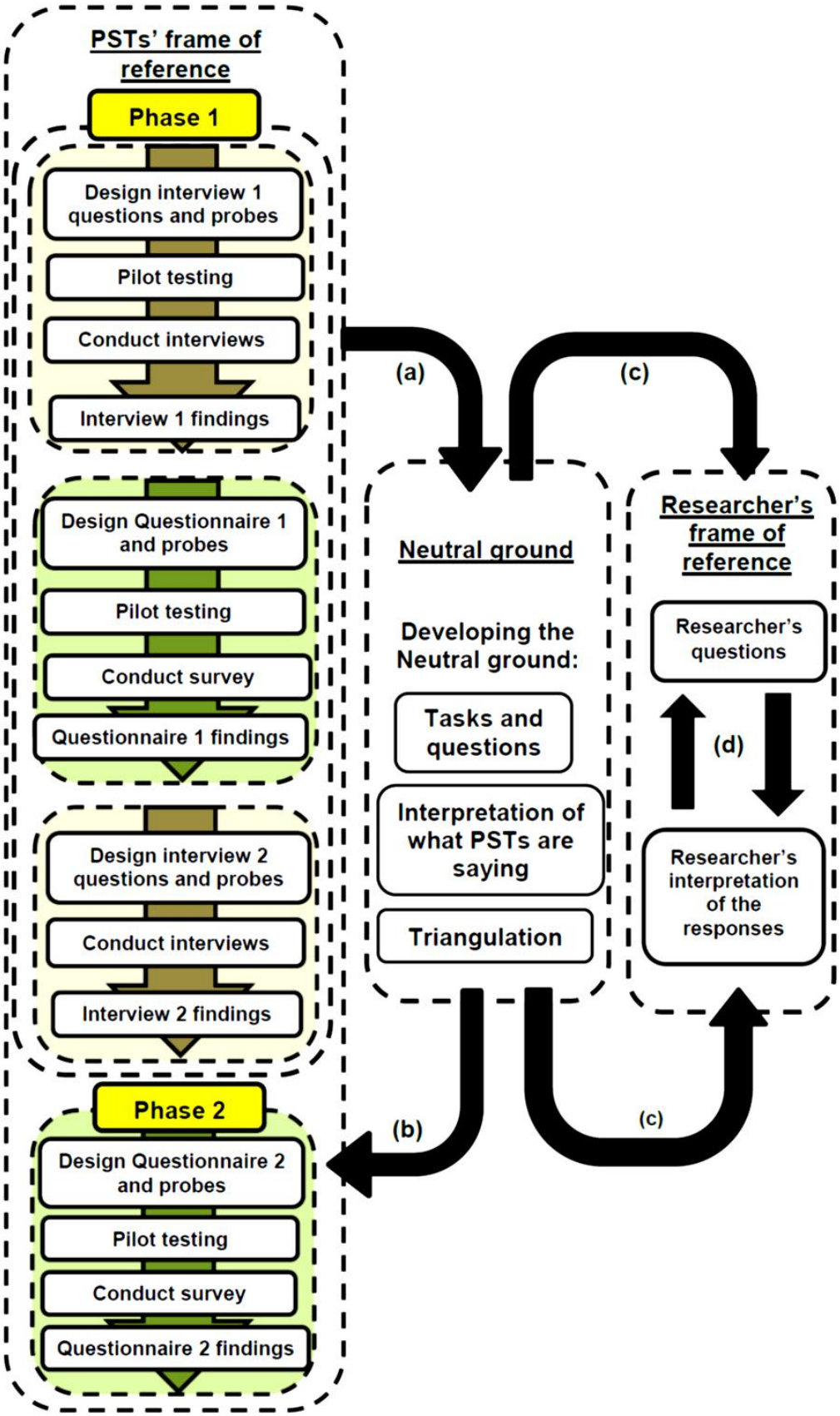
Although Chapter 3 is devoted to methodology, specific aspects like the descriptions and objectives of probes in different research instruments will be reported later in separate chapters with their results for the convenience of the reader.

3.2 Research Approach

Figure 3.1 gives the overview of the research approach, which was planned around the needs to develop a "neutral ground".

The research used a qualitative design consisting of four studies across two separate phases; Phase 1 (P1) comprised of Interview 1, Questionnaire 1 and Interview 2 whereas Phase 2 (P2) consisted of Questionnaire 2 only. The organisation of the research in two phases reflects the differences in the focus of the two phases, and that the research was carried out with two groups of PSTs at different time periods.

Figure 3.1 Overview of Research Approach



P1 was mainly concerned with Research Aim 1 (see Section 1.7) where qualitative methods of interviews and questionnaire were used to explore the PSTs' understandings of uncertainty and their procedural ideas in handling the concept. The knowledge gained from the P1 studies was then applied to P2, which was mainly concerned with Research Aim 2 (see Section 1.7) and aimed at developing a questionnaire that could be used to determine the PSTs' understanding of uncertainty in measurements.

As shown in Figure 3.1, the four studies in both phases were carried out sequentially; the data from a preceding study would be analysed to develop understanding, and to create probes or modify existing ones for the following study in order to pursue emerging ideas/concepts on uncertainty in more depth or to refine the researcher's interpretation of evidence. It must be noted the "cells" in Figure 3.1 have a dotted outline to symbolise the "porosity" to knowledge that is constantly being constructed from the gathered evidence.

Based on Figure 3.1, the research began with an interview study as this allowed the determination of the extent of research problem (see Section 1.5). The specific aims of each study in both phases will be dealt together with their results in separate chapters later. The next section explains the rationale for the research approach shown in Figure 3.1.

3.2.1 Rationale for research method: developing the "neutral ground"

The research method was built on three basic methodological principles that Johnson and Gott (1996) had identified earlier in Section 2.3 for developing the "neutral ground" (see details in the "Neutral ground" in Figure 3.1).

The P1 studies were conducted sequentially so that findings from one study could inform the planning and design of the instruments in subsequent studies (arrows (a) and (d) in Figure 3.1). Additionally, the order in which

different study methodologies had been deployed (i.e., interviews alternating with questionnaires) allowed clarifications to be made and inferences to be drawn on the efficacy of the probes. All these led to the probes being refined and the creation of new probes (arrow (b) in Figure 3.1) in the subsequent instruments. Johnson and Gott emphasised that the iterative process of finding evidence in order to refine interpretations and develop "neutral" questions increased the validity and reliability of the research findings.

Other steps were also taken to achieve the development of a "neutral ground". For instance, the chosen probes were set in the primary science context, one in which the PSTs were very familiar with. The questions in the probes were carefully crafted for easy understanding of intent; scientific jargon and ambiguity were avoided so that the PSTs' thinking would not be impeded. Pilot studies were also planned to help in the process of developing suitable probes.

Importantly, the triangulation process was well-integrated in the research approach. The assessment of the PSTs' procedural ideas using different probes in different contexts and research instruments helped to reveal the PSTs' patterns of understanding and increased the reliability of evidence. Essentially, the strategy helped to expose the multiplicity of understandings that the PSTs might have with respect to procedural concepts. At all times, I was cognisant and mindful that my interpretation of the PSTs' responses should be driven (arrows (a) then (c) in Figure 3.1) by the PSTs' frame of reference only, and not my own.

Additional steps were taken in efforts towards accurately interpreting the PSTs' responses, for instance, I would ask the PSTs several times in different ways during the interviews. Following each data collection, the PSTs'

responses to the research instruments were also shared with my supervisors, NIE colleagues, in-service primary teachers, and the PSTs themselves in subsequent class meetings in order to ascertain whether I had interpreted the responses correctly.

3.3 Research participants

The method of sampling in this qualitative exploratory research involved purposefully selecting the participants that best helped me to answer my research questions (Creswell, 2008). Further details about the participants in the two phases are given next.

3.3.1 P1 participants

At the point when the study was conducted, the P1 participants were attending a compulsory module, Curriculum and Pedagogy in Primary Science. The total number of participants involved in P1 was 55; 50 females and 5 males, ages between 21 and 25 years, and their profiles were not atypical of the whole cohort. All were in Year 2 of a four-year degree course that would eventually bring them to teach Primary Science and other subjects (for e.g., English Language, Mathematics) in a local primary school. As for their academic backgrounds, forty-one reported to have a GCE “A” Level and fourteen a local Polytechnic diploma in various technical courses. All took Mathematics and at least one science subject in their GCE “O” level examinations. Only three reported they did not take Mathematics or Science in their Polytechnic or A-level courses.

3.3.2 P2 participants

When P2 was conducted, the P1 participants and their cohort were not available as they were already in schools for practicum. Besides, I had already left the NIE, and therefore, contact with the PSTs was rather limited; moreover,

application through official channels for access (which was necessary by then) would have taken a long time. The access to PSTs for P2 was only gained via personal contacts with former NIE colleagues. Although the P2 participants were from a different group, the two batches of PSTs had similar profiles including academic and NIE course enrolment. Nevertheless, I viewed the situation as an opportunity to test the developed questionnaire.

Twenty PSTs in all were available for Questionnaire 2, and all were in their second year of the Diploma-in-Education programme. Their age range was between 21 and 25 years; eighteen were females around 21 to 22 years of age whereas their two male counterparts were between 23 and 25 years, the additional years in the male PSTs were due to the years spent in conscription. Their highest level academic qualification was also either a GCE "A" level (16) or Polytechnic Diploma (4); both were equivalent in terms of entry requirement to primary teacher training programmes at NIE. Singapore's Ministry of Education (MOE), the sole organisation that hires school teachers, ensured that individuals would come into pre-service teacher training with the right credentials at the point of recruitment.

3.4 Research Instruments

Two modes of data collection were used for this research: interview and questionnaire. As seen in the review of literature in Chapter 2, both had been extensively employed in studies to investigate student's understanding of uncertainty in measurements. Lubben and Millar (1996), based on their extensive work in the Procedural and Conceptual Knowledge in Science (PACKS) project, described both qualitative methods as appropriate for exploring students' understanding of evidence and would have the potential strength of delivering good research data.

3.4.1 Interviews and questionnaires for data collection

A copy of the instruments used in this research: Interview 1, Questionnaire 1, Interview 2, and Questionnaire 2 can be found in Annex 3.1, 3.2, 3.3 and 3.4 respectively.

Interviews

A semi-structured interview protocol as suggested by Drever (1995) was deployed in the two P1 interviews. This not only allowed a number of pre-determined questions to be asked in a systematic and consistent manner which was critical in deriving a pattern of PSTs' understandings of uncertainty but it also provided better time management for data collection. Although the questions were prepared beforehand, I took advantage of the open-ended nature of the format to digress and probe beyond the answers only when required to. The nuanced responses from the PSTs could also be explored by making up questions along the way but keeping the original intent in mind. Whenever a PST gave a vague response or had difficulties interpreting the interview question, I could support by paraphrasing the question or used examples and analogies (but not directing them to a particular response). In several probes, the PSTs were asked to fill tables with hypothetical data so that they could refer to them while responding (thus, preventing information overloading). In addition, the PSTs were also given a copy of the instrument and a pencil at the start of the interview, which not only allowed the PSTs to write and organise their thoughts before articulating a response but also permitted the visual learners among them to draw and map out their thoughts.

Nevertheless, there were several difficulties in the administration of the interviews; these included scheduling the PSTs and organising sessions to make up for cancellations.

Questionnaires

Questionnaires could neither be modified “on the spot” to assist PSTs with difficulties interpreting the probes nor would they allow me to clarify their nuanced responses. Understanding the impact of these problems on the reliability of evidence, there was a deliberate attempt at using short close-ended questions in the probes. These questions were also simply worded so that the PSTs were able to quickly grasp their meanings and respond accordingly. This did not mean open-ended questions were avoided; on the contrary, they were used whenever the intent was to get a range of responses, and normally, such responses were expected not to be lengthy. Further illustrations of the types of question will be given later.

Despite the few shortcomings mentioned, the questionnaires provided a quick and efficient way of obtaining research data. Besides, from my own observations, they were less stressful for the PSTs to respond to compared to the face-to-face interviews.

Each of the methods with its own strengths and weaknesses complemented when used together (Arksey & Knight, 1999); the questionnaires were used to check on the strength and incidence of the accounts that the interviews seemed to suggest, and vice-versa. Thus, together they provided a means of exploring and affirming the prevalence of different procedural ideas related to uncertainty in measurements.

The fact that the questions in both instruments were pre-determined helped to enhance the validity and reliability of the instruments in several ways (Patton, 2002). First, a complete set of data could be gathered from each PST ensuring that all understandings that must be assessed were eventually tested. Second, the approach reduced researcher’s bias since the questions were pre-

planned. Finally, since the PSTs answered basically the same questions pitched at similar cognitive level, the analysis of data was more streamlined especially during the coding process when the response data were compared to draw on common understandings.

3.4.2 The use of probes

The literature review in Section 2.3 had shown the widespread use of "probes" in ascertaining how individuals understood the concept of uncertainty in measurements. "Data probes" used in the current research had been likewise employed by Millar et al. (1994) in the PACKS project to extensively explore children's ideas about aspects of "reliability of measurement, about the use of numbers to represent physical quantities, and about logical reasoning from data to conclusions" (p.212).

According to Southerland, Smith, and Cummins (2000), probes generally would consist of a stimulus material (e.g., pictures, diagrams, or data) supported by a set of questions that were purposefully designed to elicit the individual's understandings of a specific concept. Thus, for instance during the interview, the PSTs were often asked to use their own words to explain a concept, and typically requested to articulate their understanding of uncertainty or a related concept by solving a problem or explaining an observation. Furthermore, diagrams and tabulated data were also given to help the PSTs "visualised" and processed the information before giving their responses.

Kvale's (2008) suggestion to researchers using probes in an interview was followed; he proposed that participants should be listened to closely in order to observe their response to a probe, and when necessary, to follow-up with questions that could help them clarify their initial thoughts, actions, and reasoning as well as to reveal their conceptions sufficiently. Another came from

Southerland et al. who suggested the participants' application of personal meanings in their responses should be focused so that the researcher would be able to tease out what was learned by rote with minimal knowledge from what could be regarded as meaningful understandings.

Despite the wealth of evidence the probes could provide, I would concur with Southland et al.'s observation that a significant amount of time could be spent trying to make meaning out of the response data during the analysis. Additionally, as pointed out by Kanari and Millar (2004), the participants might not have a "feel" for the quality of the given data in the probes; thus, unless the participants were given time to assimilate and evaluate the given data, they would not be able to respond to the probe effectively.

An important aspect of a probe was its context. Quite often, an investigation context would be used since it provided an authentic scenario that helped the PSTs to realistically apply their procedural concepts and respond accordingly to the stimulus presented (Leach et al., 1998). But the evidence gathered might only be peculiar to the probe, which made it difficult to generalise the understanding. So occasionally, a decontextualised probe was used to elicit general understanding; but I was also aware that such a probes could lead the PSTs to imagine different contexts which might result in varied responses that could cause difficulties when the response data were interpreted eventually. Careful attention was also paid to the probes' level of difficulty so that the PSTs' ability to express their understandings would not be impeded. All probes were deliberately set within everyday contexts that required minimal substantive knowledge. Examples of probes are shown in Figures 3.2 and 3.3.

Figure 3.2 Probe on repeated measurements in an investigation (Interview 1)

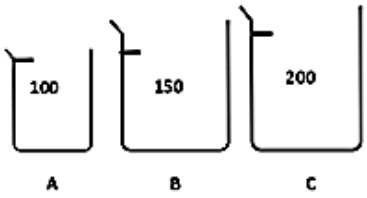
A scientist wanted to see how temperature affected osmosis in a potato. She puts equal size chips in a 1 mol/dm^3 sugar solution at different temperatures and measured the change in mass of the 'chips' on a top pan balance and recorded the results as a percentage of the original mass after 4 hours. She repeated all the readings 3 times.

Percentage of original mass			
Temperature (°C)	1	2	3
15	83.46	78.88	80.91
30	77.5	82.67	80.59
45	82.66	65.47	74.24
60	67.76	93.32	73.18

Question 9: What does the data show? Explain all your reasoning?

Figure 3.3 Probe on single measurement (Questionnaire 2)

(b)(i) Which beaker would you choose to get a volume of water of around 80 cm^3?



Circle your answer | It doesn't matter which | A | B | C

Explain your choice in (b) (i) _____

The examples can be used to illustrate some general points about the accompanying questions used in the probes. The accompanying question in Figure 3.2 was open-ended so as to allow the PSTs to synthesise their procedural ideas without being "forced into response possibilities" (Creswell, 2008, p.226). In Figure 3.3, the question began with a multiple-choice consisting of all possibilities followed by a more open-ended question requesting the PSTs to explicitly state their supporting reason; questioning this way allowed the pre-determined close-ended responses to yield useful information especially when a breakdown analysis was done later. The whole format, therefore, prevented the PSTs from merely giving a routine close-ended response without reasoning.

The source of the probes used in P1, namely Interview 1 and Questionnaire 1 (the initial studies) was from a bank of probes that were

developed over the years [see Collins Science Investigation Series written by Gott, Foulds, Johnson, Jones, & Roberts (1997, 1999)], some of which were used recently in a study on initial teacher education (ITE) students in Durham University (Glaesser, Gott, Roberts, & Cooper; 2009a and 2009b). The probes in Interview 2 and Questionnaire 2 were either modified from probes used in Interview 1 and Questionnaire 1 or created based on the issues that emerged from these studies.

3.4.3 Pilot Studies

All research instruments used in this study with the exception of Interview 2 (which was revised based on inputs from earlier instruments) were piloted before they were deployed. Details of the pilot studies are given in Table 3.1 below.

Table 3.1 Pilot studies[19]

Instrument	Number of PSTs involved	Location/s
Interview 1	10 2	Durham University NIE
Questionnaire 1	20	NIE
Questionnaire 2	2 14	Durham University NIE

The pilots were mainly to check the time to complete the instruments, the suitability of the probes, and potential difficulties (for e.g., due to the use of technical terms, ambiguity, poor grammar, illegible diagrams, insufficient information, etc.). In addition, the pilot study for Interview 1 also allowed the researcher to practise and test the computer software (Audacity) used to audio-record the interviews. After the pilots for Interview 1 and Questionnaire 1, several modifications were made especially with regards to the accompanying questions (for e.g., more follow-up questions were added to the interview probes), the diagrams (for e.g., clearer diagrams with labels were added to help

[19] The PSTs in the pilot studies were not involved in any of the revised instruments subsequently.

the PSTs visualise the experimental set-up), and the choice of words and terms (for e.g., “trainers” was dropped for “running shoes”). The questions were also given a more local cultural “feel”; for example, English names of fictitious investigators were replaced with local ones. Extraneous words deemed irrelevant to the questions were also removed to prevent information overloading.

Questionnaire 2, the final instrument, was piloted in Durham University and NIE with pre-service primary teachers but the PSTs did not report back any problems. Nevertheless, some modifications similar to those stated earlier for Questionnaire 1 were made to the probes before the instrument was administered.

Additional steps were also taken for Questionnaire 1 to gain more insights for the development of the final Questionnaire 2. Questionnaire 1 was given to an expert panel of five NIE lecturers and five primary science teachers (with more than three years of experience) for validation (a copy of the invitation letter and validation feedback can be found in Annex 3.5 and 3.6 respectively). The feedback received was quite minor; they were mainly concerned with the wordings and terms used in the probes. Some panel members suggested changes to the presentation format to make it easier for the PSTs to retrieve information from the probe. All feedback and suggestions were carefully considered and appropriate changes were then made.

3.5 Data Collection

The time available to study the P1 participants was about a semester. Since the period was quite short, the data collection became quite intensive. However, it also meant the participants were unlikely to change much in terms of their understanding of uncertainty, and this was critical to the success of the

triangulation strategy mentioned earlier. Nonetheless, the PSTs were monitored to see if they had received any instruction related to the topic of the research during the period, but none actually did.

Interview 1 and 2

Interview 1 was conducted with twenty-eight PSTs and Interview 2 with the remaining twenty-seven five months later. There were no specific criteria for selecting the interviewees. The PSTs were interviewed based on a schedule drawn to fit their free time. Each interview lasted between 35 and 45 minutes and was audio-recorded only. The duration did not include the pre-amble. The latter was used to tell the interviewees to be calm and to give their best answers; the purpose of the interview for academic research was reiterated again to allay fears, and the PSTs were also informed about the confidentiality of their responses.

At several points during the interviews for certain probes, the PSTs were requested to "think aloud". This allowed me to elicit the "inner thoughts or cognitive processes that illuminate what's going on in a person's head during the performance of a task" (Patton, 2002; p.385). Good practices suggested by Kvale (2008) were observed: the interview was conducted in a relaxed manner; starting and completing on time; and thanking the interviewee at the end of the session. Samples of transcripts from Interview 1 and Interview 2 can be found in Annex 3.7 and 3.8 respectively.

Questionnaire 1 and 2

Questionnaire 1 was administered and completed in one seating for all fifty-five PSTs. Questionnaire 2 in P2 was distributed to twenty PSTs at a much later date and completed in one session as well. A time of 1 hour was allocated for the completion of both questionnaires, which was more than enough (based

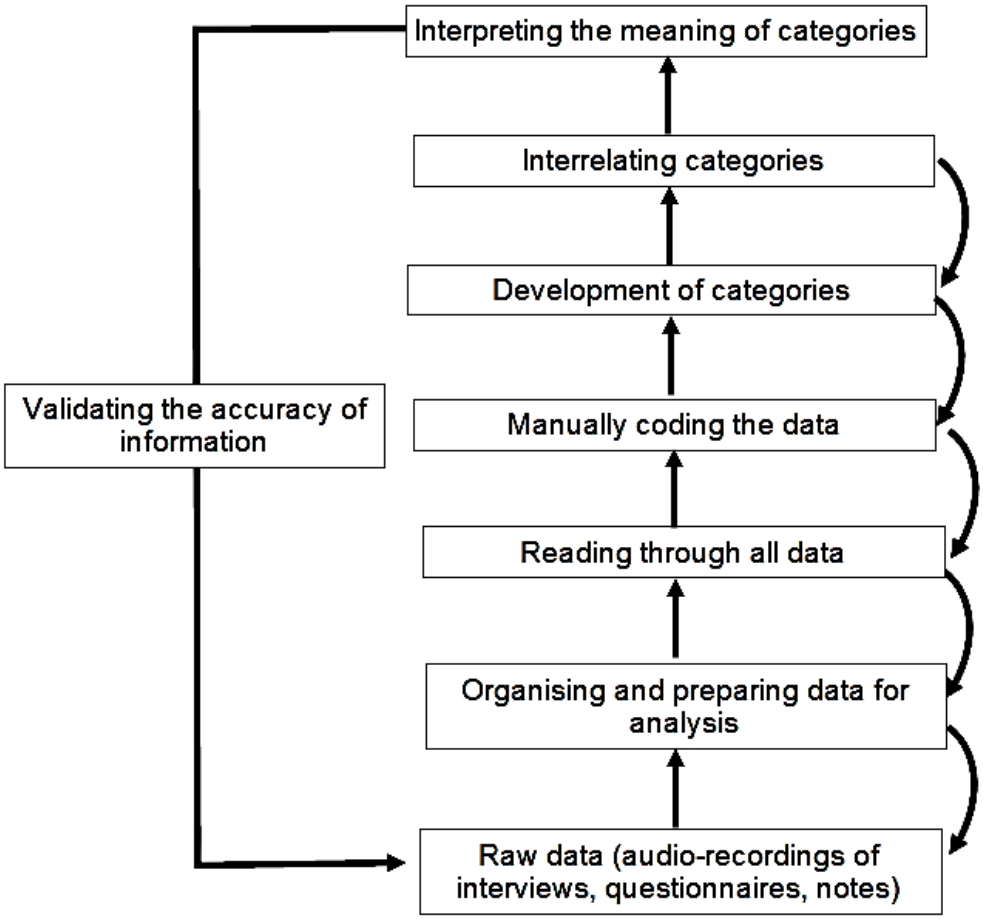
on the pilots). Like the interviews, before the start of the questionnaire, the PSTs were informed about the research purpose of the questionnaire and asked about their mental and emotional conditions. The researcher was present throughout the session to answer queries but no questions were asked on both occasions when the questionnaires were administered. Samples of completed questionnaire from Questionnaire 1 and Questionnaire 2 can be found in Annex 3.9 and 3.10 respectively.

3.6 Data reduction and analysis

Coffey and Atkinson (1996) described qualitative analysis as an eclectic process and cited Tesch (1990) who believed data analysis should be flexible since there were no rules to guide how it should best be done. Coffey and Atkinson also claimed there was no "one correct approach or set of right techniques" to analyse qualitative data and suggested the process should be "imaginative, artful, flexible, and reflexive", but at the same time, "methodical, scholarly, and intellectually rigorous" (p.10). The researchers further suggested that analysis in any qualitative study must basically comprise two stages: the first would involve several tasks like "coding, indexing, sorting, retrieving, or manipulating the data", and second, would essentially be the "imaginative work of interpretation" (p.6) that not only highlighted the unique characteristics in the evidence but also the "regularities of incidence or pattern" (p.7) of the participants' responses.

Bearing the preceding points in mind, a modified approach based on the Grounded Theory (Strauss & Corbin, 1990) was used to analyse the qualitative data. Different researchers have shown how this could be done. The analytical procedure to analyse the research data in both the P1 and 2 studies followed a method proposed by Creswell (2009) (see Figure 3.4).

Figure 3.4 Steps for analysing the research data (modified from Creswell, 2009)



As shown in Figure 3.4, the derivation of evidence was done inductively, going from detailed data to the general codes, and finally the categories. It is important to emphasise that Figure 3.4 does not intend to suggest the analysis was carried out in a linear and hierarchical manner; rather it was done iteratively with the various stages interrelated and not always visited as presented. The raw data, for example, the audio-recordings from the interviews were first examined to see whether the interviews were fully recorded and there were no missing parts. Incomplete or "spoilt" recordings were discarded (only two such cases in Interview 2). As for the questionnaires, missing answers to the probes were generally accepted as non-response, and the PSTs were assumed not to know the answer. The raw data was then prepared for the next stage of the analysis.

The interview recordings in this research were transcribed using guidelines proposed by Creswell (2008) who suggested noting the interviewee's reactions (for e.g., facial expressions, head nodding, and laughter) or major interruptions (for e.g., a phone call) as all these could help get a better sense of the nuances in the PSTs' responses. In transcribing the interviews, additional information was also retrieved from notes written during the interview (could be about the interviewee's interesting "body language") and the PSTs' scribbling on their copy of the instruments.

Both transcripts and questionnaires were arranged according to the order in the class list obtained from the Registrar, and the names of the PSTs were then permanently replaced by a unique simple code (an alphanumeric) that allowed easy retrieval of the scripts, and more importantly, to identify the source of evidence (especially for tracing the author of an extracted piece of data).

The transcripts or questionnaires were skimmed across several times to get a general sense of information they contained and to reflect on the PSTs' overall understanding of concepts. The second reading was done slowly in two stages. First, each transcript or questionnaire was read in order to get a holistic description of the PST's thinking; referring to questions that were conceptually related helped in the understanding and sorting out of responses that appeared to be unclear or ambiguous. The second reading focused on responses to a particular probe; this was done to get the range of perspectives the PSTs held when they dealt with a particular probe, and specifically to the concept that was tested. Reading the responses across different individuals would reveal patterns of understanding for the whole group of PSTs.

After reading, next came “data reduction” where the large amount of raw data was reduced into analysable units (Coffey & Atkinson, 1996). For convenience, Patton (2002) suggested the responses should be organised by their initial question. Thus, my first action was to segment all responses from a single transcript or questionnaire by probes. To be able to view and make comparisons of *all* data (for each probe), the responses were placed in a single “basket” (a Microsoft Excel worksheet). At this juncture, the response data were still “untouched” (however, using the functions provided by Excel, a few would carry sense-making comments from the second reading).

Following the organisation of response data by probes, the next stage was the detailed process of coding; in this aspect, Creswell (2009) described coding as a process of organising the data into “chunks or segments of text before bringing meaning to the information” (p.186). In the coding process, the long raw responses would be condensed into fewer words or a phrase (*usually words or phrases used commonly by the participants*) but still retaining the main sense of the response. Each phrase that had a different meaning would be recorded separately. The phrases would be compared and contrasted with those already tabulated in order to look for convergence in meaning. Patton (2002) described this as looking for “recurring regularities” (p.465) in the data. There was also a constant review of the coded outcomes in order to reduce overlapping or redundant codes. The process of clustering the phrases would continue until all the responses to a particular probe were examined and sorted into categories.

The derived categories would represent the whole range of understandings to a particular probe. For instance, in Figure 3.3, PSTs would respond to the probe by supporting “**A**” or “**B**” or “**C**” (which represent different

capacity beakers) or “it does not matter which”. They would have to justify their choice by giving a reason, which would then be analysed and categorised (if there were more than one reason for a particular choice).

How were the categories used? By calculating the frequency of similar codified phrases for each category, quantitative results were obtained to reflect the PSTs’ predominant procedural ideas. Sometimes, interrelating the categories of probes that tested a similar concept of evidence led to bigger categories. For example, in Questionnaire 2, there were a series of probes that looked at the PSTs’ choice of instrument to take a single measurement (see Figure 3.3). The PSTs had to choose the best instrument (for e.g. a thermometer) from a range that differed only by a single feature (for e.g., range of scale, resolution, etc.) to measure a given quantity accurately. The categorisation of coded responses to the different probes allowed us to see the PSTs’ use of concepts (viz. the resolution of scale and the concept of full-scale deflection) in deciding their choices. Finally, the derived categories of PSTs’ understandings of a particular concept allowed comparisons to be made across different probes in the same instrument or the same probes across different instruments.

It is important to emphasise the process from raw data to developing categories (see Figure 3.4) was carried out iteratively; there was constant shuttling back and forth between different stages to check meticulously if the analysed data were accurately described and interpreted. This essentially addressed “interpretive validity” (Miles & Huberman, 2002), which was critical for this study. Bearing the same notion in mind, to get a wider feedback on my data analysis, methods and findings, the study was also presented at two

international conferences[20]. Although no specific comments were given for the data analysis, I obtained positive and motivating feedback for my findings.

3.7 Actions taken to address ethical issues

Before the studies began, the following actions were taken:

(a) The Head, Natural Sciences and Science Education (NSSE) as well as the Deputy Head (NSSE), who was in charge of curriculum matters, were informed of my intentions to conduct a doctoral research with PSTs attending modules conducted by the academic group. A verbal approval was given to me to proceed with my research on the condition that all personal information divulged by the participants was kept confidential and the PSTs' participation in the study was purely voluntary.

(b) The PSTs were informed of the purpose of the research and the modes of investigation conducted during the study. The PSTs were told that their participation was purely voluntary, their identity plus the information they divulged would be kept strictly confidential and would only be used for research purposes. In addition, the PSTs were not pressured into giving answers and were assured that their performance in the research would not have any bearing on their course grades.

[20] 1. 41st Annual ASERA Conference (Australasian Science Education Research Association) (Port Stephens, New South Wales, Australia, July 2010); 2.The 2nd East Asian International Conference on Teacher Education Research (Hong Kong, December 2010)

3.8 Summary of Chapter

The chapter described the rationale behind the research approach, which consisted of studies in two phases. The efforts geared towards getting valid and reliable interpretations of the PSTs' understanding of uncertainty in measurement entailed a methodology that sought to develop "neutral" instruments through a triangulation strategy that tapped on different probes and research instruments.

In addition the chapter presented reasons for its data collection methods, the use of probes, and the kinds of questions that tested the PSTs' understanding of procedural concepts. It provided details of the research participants as well as the measures taken to address ethical issues. The chapter also explained how the analytical process of developing evidence for this research; the iterative process of reducing the data by coding and arriving at categories where some quantification of results could be obtained helped in enhancing the reliability of its research evidence.

CHAPTER 4

PHASE 1 INTERVIEW 1 STUDY

4.1 Chapter Overview

This chapter concerns Interview 1 (P1I1)[21], the first of three studies in Phase 1 of this research. It will first state several major aspects of the instrument and then introduce its key aims. This will be followed by the main section, which consists of the description of each probe that includes its specific objectives, results and discussions, and a review of the probe. A copy of the P1I1 interview protocol and a transcript from the interview can be found in Annex 3.1 and 3.7 respectively.

4.2 Structure of P1I1 Interview protocol

The P1I1 instrument consisted of four probes; all focused on repeated measurements, and as shown below:

Probe 1: procedural ideas about repeats

Probes 2 and 3: uncertainty in a set of repeated data (Type 1 investigation)

Probe 4: uncertainty in different sets of repeated data (Type 2 investigation)

The face-to-face interview study involved twenty-eight participants.

4.3 Main aims of P1I1

(a) Being the first, P1I1 served to reaffirm initial observations of the PSTs' understandings (or the lack of it) of uncertainty in measurements (see Section 1.5).

(b) The P1I1 instrument was developed mainly to focus on Research Aim 1. It was designed to investigate different conceptions PSTs might use in understanding uncertainty by probing the "thinking behind doing" repeated readings. The different conceptions include the purpose of

[21] For convenience, "Interview 1" will be referred to as "P1I1" which stands for "Phase 1, Interview 1". Subsequent studies will be abbreviated the same way.

repeats, the causes of uncertainty, the sources of errors, the existence of true value, and the idea of a "perfect" instrument or method of measurement. The study was also developed to surface the PSTs' procedural ideas related to uncertainty in situations that required them to predict or interpret data in investigations or to propose plans as a way of dealing with the uncertainties and moving forward in their investigations.

4.4 P1I1 probes: objectives; results and discussions; and review

4.4.1 Probe 1

> Scientists usually repeat readings if possible, rather than just taking one. Why do they repeat readings?

Specific objectives

Probe 1 looked at fundamental idea/s that motivates the PSTs to repeat a measurement in scientific investigations. The term "scientists" was intentionally used here because they were assumed to be standard-bearers of good scientific inquiry; the statement therefore implicitly suggested that getting accurate and precise measurements (quality results) would be an ultimate goal.

With respect to its objectives, the responses would help explore the PSTs' purposes of taking repeats, whether they understand that repeats are meant to "capture" uncertainty (or variation) in measurements, or whether it is a way of finding a true value (as reported in the literature), and if so, how this idea of true value will be manifested.

Results and Discussions

After coding analysis, the PSTs basically gave five ideas why scientists repeat their measurements:

(a) For accuracy(11);

(b) A checking strategy(8);

(c) A routine practice(2);

(d) To find a mean(5);

(e) To get a spread/range (2).

It is important to note the PSTs usually provided one and occasionally two ideas, which explained why the total number of ideas (shown within parentheses) was above the number of interviewees. The PSTs' responses were generally short and lacking in details, and this was indeed how most responses in this research were given as well. To improve the responses, the PSTs were posed with follow-up questions or by asking them to elaborate their initial response.

The term "accuracy" was often given spontaneously but it seemed to have a rather "loose" meaning. Frequently, it meant a set of consistent[22] values that some suggested could be used to obtain a mean value. Their descriptions did not relate to "accuracy" as described in Section 2.2.2, rather to the concept of "precision". Additionally, a number of responses seemed to suggest the notion of true value; a number of PSTs in the group seemed to think that a reading which appeared several times was unlikely a random phenomenon but could possibly be the one true value.

Interestingly, the PSTs had different ways of recognising "consistent" data; for instance, some claimed it should be a value that appeared a number of

[22] "Consistent values" could mean precise readings whose small differences might not be recorded because of the limitations of the measuring instrument but such thinking was not shown in any response.

times (e.g. three) *consecutive*ly but others were contented with the same value appearing *scattered* within a data set, as seen in the response from I1-20[23]:

> I think they want to find out the most accurate result so they repeat the readings …maybe the second and the fourth are the ones that tally, if they kept getting this result repeated times, maybe this is the most accurate one. (I1-20)

In the second category, "a checking strategy", some PSTs claimed repeating measurement would allow scientists to check and replace one or more readings:

> Because when you do the experiment, the results may never be the *same*…there may be one or two readings that are out of range, so in order to replace them, you have to repeat for that one or two readings. (I1-9)

The third category had only two PSTs, and they held the idea that repeating a measurement was an established laboratory routine:

> Everyone does a few readings so it's like a culture to find [one]. (I1-12)

In the fourth category "to find a mean", there were PSTs who demonstrated conceptual understanding of variation by relating it to unpredictable "errors" or "factors" with environmental conditions being frequently mentioned as an example. They favoured the measure of central tendency because they believed it balanced out the fluctuations in the repeated values caused by the errors/factors:

> There are errors…sometimes they overlook other things [uncontrolled variables], so it's like taking the average out of everything. (I1-13)

The mean value obtained from a set of repeats therefore was held as the best representative for a data set:

> We take an average as this reduces the amount of errors caused by both humans and instruments. (I1-18)

But this very same notion could lead others like I1-26 to claim the mean was in fact the true value:

> If you take a lot of readings, let's say N times, you get to see whether there is a true value in the experiment, [which] in this case, is the average. (I1-26)

[23] "I1-20" is an example of the alphanumeric code used to identify the PSTs in this research; it stands for the twentieth PST in the list of participants. The code would be unique to each PST.

In the entire group, only two PSTs described repeating measurements as a way of "capturing" the spread in data. Both PSTs also conveyed that variation could be caused by uncontrolled factors; the response from I1-9 below conveyed this:

> Normally they [scientists] give a range…There are always errors or [uncontrolled] factors that would influence the experiment. (I1-9)

On the question whether true value existed, it was brought up with *only* eleven individuals whose initial response suggested it was their aim in repeating measurements. Eventually, only five (about 18%) PSTs claimed it existed, and finding one was indeed their purpose of repeating a measurement:

> I think there's a true value and we try to [get] as close to it as possible. (11-8)

> Experiments are [meant] to get the true value or how to get to a true value. (I1-4)

In addition, among the eleven PSTs, two seemed to bear the idea that *all* values reported in textbooks or handbooks were indeed true values (i.e., error-free) that they could obtain in their measurements.

The remaining PSTs who claimed true values did not exist highlighted that measurements would always be affected by errors from measuring instruments, human investigators, and fluctuating environmental conditions.

Review of Probe 1

The lack of experimental details in the question led to some initial confusion; repeats were wrongly assumed to be a single investigation performed several times or a set of DV readings taken for a range of IV values. Thus, during the interview, the idea of repeated readings was explained and the PSTs' interpretation of the probe was regularly checked.

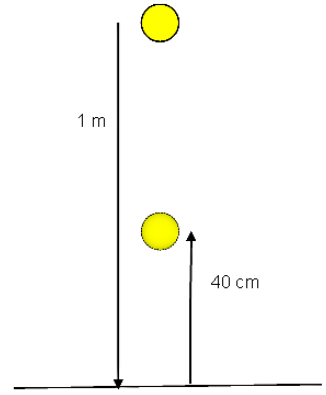
The term "accuracy" was frequently used by the PSTs as compared to "precision"; nevertheless, many did not seem to be able to articulate their meanings or distinguish them. Since their understandings were crucial to the

overall concept of uncertainty, it would be essential to clarify the PSTs' conceptual understanding of these terms so that the interpretation of their responses could be done accurately. The problem after all could be due to the lack of vocabulary rather than poor conceptual understanding. This will be investigated in other P1 studies.

The idea of true value was surfaced by the PSTs without being prompted by planned questions. The probe showed that about 18% (5) might have believed it was not hypothetical but a real value, one that could be obtained through measurements (and the chances of obtaining one increases by repeating a measurement). As shown by the evidence, true value could be expressed using different terms (for e.g. "exact", "correct", "right", etc.). Thus, it would be important to carefully interpret the PSTs' responses in order to establish what they really meant.

4.4.2 Probe 2

Imagine a squash ball was dropped from the height of one metre and you measured its rebound height against the metre rule. It bounced back to 40.0 cm.



Bounce height in cm										
40.0										

The accompanying questions will be given separately

Specific objectives

Probe 2 was a familiar primary-level investigation low in procedural complexity. It consisted of a single dependent continuous variable (rebound

height) responding to a single independent categoric variable (squash ball). This was deliberate so as not to impede the PSTs' thinking and response to the five accompanying questions.

The expected responses to the probe would help answer several key questions: Did the PSTs understand that uncertainty will *always* be present in measurement? What were the PSTs' causes of variation in repeats, did the PSTs put it all down to human errors, or did they know other sources of random errors such as uncontrolled variables? Could the PSTs relate the degree of variation to the number of readings or the reliability of measuring instruments? Finally, did the PSTs have the notion of a "perfect" method of measurement or instrument?

Results and Discussions

Question 2:

> If you were to drop the ball again, what reading do you think you'd get? Fill in the next space in the table.

For the second reading in the table, instead of giving one value, fourteen PSTs stated a range between 38 and 42cm, and added it might also be 40cm (which was given as the first rebound height). Nine gave a single value other than 40cm but close to it, for instance, 39 or 41cm. The responses from all twenty-three individuals indicated they expected to see variation in the data. The response from I1-13 below illustrates this:

> This is what we call experiment; we can't get very similar answers…there are variations. If it's fixed [the same value] I think there's something wrong with the experiment. (I1-13)

The twenty-three PSTs were asked to account for the difference between the first and second readings, but only twenty responded with some giving more than one reason. All the different ideas were pooled together, coded, and then categorised to see the patterns of understanding. The

following categories were obtained, and the numbers within the parentheses would represent their frequencies:

(a) Human errors, for instance, a force was exerted on the ball when released(7), errors due to parallax error or inconsistent way of taking maximum heights(5);

(b) Uncontrolled variable, for instance, effects of draught(6), the ball losing its elasticity(1), and the spot where the ball landed was uneven(2);

(c) Theoretical reasoning that errors were bound to occur (2).

The distribution showed that human errors (a) were often cited as the cause of variation. In fact, of the twenty PSTs, seven actually cited human errors as the *only* cause of variation.

Among the twenty-eight PSTs, four indicated the second height must be 40cm (similar to the first height) and could not be anything else suggesting they did not expect uncertainty at all. Three of them explained the height should not change since the ball was released from the same height for every repeat; one of them added another reason:

> Because…it's me [the same individual] who did it the second time, it must be the same. (I1-10)

The fourth PST (I1-8) based her prediction on a misconception; she assumed "the P.E. and the K.E. will be the same" for all repeats, and therefore, "the distance the ball bounced to will be 40cm".

Question 3:

> Imagine you bounced it 10 times. What would the other 8 readings look like? Fill in imaginary results into the table.

The majority (24) described the next eight heights should be close to 40cm; they used phrases such as "35 to 45cm" or "40±2cm" or "slightly more or less than 40" or "±5cm". The PSTs were asked if 40cm could be repeated in

any of the eight rebound heights to test their probabilistic thinking. All except one said it was possible; the exceptional PST claimed her intuition led her to think the eight readings should *always* be higher than 40cm:

> To me I think it will be above 40…actually I don't really know how to explain this. Based on my intuition…it is unlikely to fall below 40. (I1-24)

I1-24 could be thinking about systematic errors and was confused over the contribution of this error to the height measurements.

As a triangulation strategy, the cause of variation was asked again to see if the PSTs had changed their earlier ideas. Seven PSTs did not respond; and the responses from the remaining seventeen can be distributed as follows:

(a) Human errors, for instance, a force was inadvertently exerted on the ball when it was released(5), errors due to parallax error or reading error when taking height measurements against a ruler(9);

(b) Uncontrolled variable, for instance, effects of draught(7), the spot where the ball landed was uneven(3);

(c) Theoretical reasoning (1).

Like the previous question, human errors (a) were once again cited as the main cause of variation. The number of PSTs who gave human errors as the *only* cause however was reduced to four and only two PSTs who previously held such a view in Question 2 remained.

Five PSTs claimed the heights should remain the same throughout all ten readings, and this was the same PSTs who claimed the second height should be the same as the first in Question 2. They rationalised by again citing the fixed height from which the ball was released; the response below underscores this point:

> It should be around 40cm because it's the same height [from where the ball is released] all the way. (I1-20)

Question 4:

What about 50 or 100 times? Explain.

Eleven PSTs claimed they would still get a similar range of values as the first ten readings although the number of repeats was increased to 50 or 100 times. According to them, this was because the experimental conditions were not altered:

> It should be around the same unless there are external factors that affect the readings. There would not be much variation. (I1-9)

Five predicted a bigger fluctuation of values due to investigator's fatigue that resulted in the ball being released from different heights unknowingly:

> It depends on the person who dropped the ball. If the person is tired of throwing then it will be more or less than the normal. (I1-6)

A sixth PST from the same category claimed the bigger spread could be due to changes at the spot where the ball had landed many times.

Two PSTs claimed there would be a certain pattern in the rebound heights, for instance, I1-24: "Initially it will be more than 40, then slowly below 40". The two PSTs expected to see a particular pattern of repeated measurements to emerge and such conceptions might have come from their past learning experiences:

> I have done this a few times, every time the results are similar; the first results were higher but the subsequent results tended to be lower. (I1-4)

They might also be referring to a well-established routine of carrying out preliminary trials (see Section 2.5.3) to gauge the range of values for the DV repeats. As a form of best practice, local students including the PSTs were often instructed to get "rough data" first, which could then be used to set the upper limit of the measurement. This might lead some into assuming that readings taken after preliminary trials would always be lower.

Another category of four PSTs indicated their readings would remain the same throughout or at least some sections of it because of the "practice effect":

> The more you drop, the more constants you get, you may get 40 40 40 or 39 39 39. (I1-1)

Such a notion might have also developed from the idea of "preliminary trials", which was to allow the investigator to get a better "feel" of the measuring instruments. As for the five PSTs who earlier asserted their readings would remain 40cm for the first ten readings, they insisted the reading would still be 40cm because the experimental conditions such as the height from where the ball was released had not been altered. Two of the PSTs used their knowledge of the law of conservation of energy to justify their answer; one is shown below:

> Yes, even after 50 or 100 times, there won't be any change as long as the initial P.E. is the same. (I1-8)

When the number of readings was increased to 50 or 100 times, the intended response was expected to include ideas of getting a better representative set of measurements or a more "trusted" mean value (all derived from the concept of standard error; see Section 2.2.6). None of the PSTs related these ideas. Perhaps, there were insufficient prompts in the question to elicit the intended response or it might well be the PSTs were unaware of statistical concepts like standard error. The *complete* absence of responses seemed to be pointing the latter.

Question 5:

If instead of measuring the rebound height by eye against the ruler you'd used a video and had 'freeze framed' it at the highest point, would you have got the same results (as the 10 in the table)?

The majority (68%) felt the rebound heights could be measured more precisely using a video with freeze-frame features; thus, a common perception

was that the repeated readings would still be around 40cm but the range would be *narrower* than the one obtained using the eyes only:

> There'll be smaller range because like when you look through a video and observe through it, you are able to focus just at the point. (I1-2)

The PSTs in this category generally believed the video would reduce human errors such as parallax error or reading error in taking height measurements using the metre rule. The response was expected given the majority had highlighted human errors as the main cause of variation earlier in Questions 2 and 3.

Of the twenty-eight PSTs, only four felt the video would not have any impact on the quality of data collected, and as such, the data would have the same degree of variation as obtained using the eyes only. They reasoned by citing errors that might be caused by the human operator of the video. For them, all the advantages gained from using the video would be diminished by the introduction of *new* human errors that emerged from having to operate the video. One PST I1-24 was rather unique in her response as she claimed using the video would in fact reduce precision because its use would only compound the number of errors.

Finally, there were five interviewees who claimed the repeats would all be the *same* value as the video, being highly sophisticated, would give *very* precise readings:

> I think it will be the same because you are able to freeze the frame so that the results will be captured...The first and the second result will be the same. It would be the same throughout if I do it several times. (I1-20)

Two PSTs from this category were flagged in Questions 2 and 3 for claiming invariance in repeated readings. The evidence thus far seemed to indicate that such PSTs might also think of human errors as the *only* cause of variation, and therefore, precise readings could be obtained by addressing those errors. It is

quite likely that these PSTs might have thought of human errors as “mistakes” that could be eliminated conscientiously.

Question 6:

Do you think there is such a thing as a ‘perfect’ measuring instrument? What would a ‘perfect’ measuring instrument be like?

An essential point of clarification was needed by most PSTs before they responded to the question, and this was whether the instrument was fully automated or required a human-operator. They were told they could imagine any of these conditions but this reflected a tendency among some PSTs to assume that uncertainties were mainly caused by human errors.

After data reduction, the first category of seven PSTs claimed a “perfect” instrument could exist but three respondents including the one below believed this would only be possible in the future:

> I think there will be in the *future*. For now, no; I think there will be 99.9% accuracy but there will be 0.01% of error. (I1-9)

The members from this category described their idea of “perfect” instrument as one that was largely free of human errors implying an “idealistic” instrument that was self-automated; the response from I1-8 illustrates this:

> Let’s say the perfect measuring instrument can operate on its own; will it give the same reading repeatedly? It will! (I1-8)

Cross-checking earlier results, four PSTs from this category were identified several times before for claiming true values existed and human errors as being mainly responsible for variation. A second category of four PSTs believed that by default, measuring instruments could never be “perfect” because they were always handled by human operators who were bound to produce errors.

The final category of seventeen PSTs was unequivocal about the non-existence of a “perfect” measuring instrument as they claimed uncertainties were unavoidable:

> Every instrument is bound to have some errors. (I1-14)
>
> Every instrument carries an error, just a matter of how large it is. (I1-18)

There were also several interesting responses that reflected the strong impact of learned routines. One example is given below:

> All this while from primary school until now, I have never had a perfect measuring instrument, there's always a question: why do you think there are errors, and they don't ask why there are no errors? (I1-9)

Review of Probe 2

Probe 2 was highly successful in drawing out a variety of ideas concerning uncertainty from the PSTs. There was a slight confusion initially in Question 3 when several PSTs assumed the question referred to successive heights of a bouncing ball after a single release. Therefore, I took the initiative to clarify this at the start for about half the group. Most PSTs also skipped writing data in the table and chose instead to verbalise their predicted data. Apart from providing written evidence of variation, the table would have been useful in triangulating the findings from Probe 1. Besides, some PSTs who were not able to articulate their ideas could have expressed them better by referring to the table. Bearing all these in mind, the PSTs would be strongly encouraged to fill the table if the probe was used again.

The PSTs' ideas about the number of repeats were not probed (this will be done in another P1 study) although on hindsight, the probe would have allowed for it. Nonetheless, in Question 4, the responses indicated the PSTs' decisions on the number of repeats might be driven by pragmatic considerations (for e.g. time) or learned routines like "three or five repeats would be enough" or "the more readings, the better". These interpretations will be further investigated in another study

Compared to Probe 1, there was a decrease in the number of PSTs who expected consistent data in their repeats (see Questions 3, 4 and 5). Perhaps this could be due to the current probe being related to an investigative setting unlike Probe 1 which adopted a generalised context. The experimental context in Probe 2 would engender the PSTs to draw on their investigative experiences including their past encounters with variation in data. The context of a probe therefore needs to be considered when interpreting the response data because the research evidence seemed to imply its influence on the generation of ideas in the PSTs.

Evidence from the analysis showed that human errors were the main cause of variation for most PSTs but again this might be dependent on the experimental context. Nevertheless, it would be interesting to see how the PSTs would react if the human factor was completely removed. Will they expect variation? What about the number of repeats, will there be less repeats or just one measurement? Investigation into these questions will be done in subsequent P1 studies.

To some extent, the introduction of the video with freeze-frame features in Question 4 provided answers to some of these questions, but its more specific intent was to check whether the PSTs understood a reliable instrument would improve precision. Although the majority conveyed this, several PSTs missed the point because they overemphasised the contribution of errors by the investigator operating the video. Thus, to make the probe more effective in meeting its objective, perhaps it should be modified to include terms like “self-automated” or phrases like “operated by robots”.

Finally, we recognised that uncontrollable factors could be a major source of uncertainty in measurements. But in Probe 2, there was little mention

of uncontrolled variables as a cause of uncertainty other than citing environmental factors (draught was always used as an example). From my own experience dealing with school practical, environmental factors were often cited as a source of experimental errors, and therefore, the PSTs' response might well be a routine one. It could also be that the probe did not provide sufficient scope to mention other uncontrolled variables; however, there were a few instances in which references had been made to "the spot where the ball landed" or "the ball was unconsciously "thrown", and at different angles" or "the elasticity of the ball". The next Probe 3 shall throw more light on this issue.

4.4.3 Probe 3

A class of 12 pupils did an osmosis experiment. All did the same experiment, using both apple and potato, and pooled/collated their results on the board. They placed potato and apple 'chips' of equal size and mass in a sugar solution and measured the change in mass of the 'chips' on a top pan balance and recorded the results as a percentage of the original mass after 4 hours.

Their results are in the table below:

	Percentage of original mass											
Apple	105.03	104.49	107.38	104.69	107.36	105.63	105.25	104.37	102.97	99.02	104.69	104.77
Potato	95.24	95.31	94.00	99.43	95.17	93.42	94.73	94.02	93.96	101.07	96.83	93.31

Question 7: Why did the 12 readings for 'apple' differ from each other?

Question 8: Imagine if you'd been able to ensure that the pupils did EXACTLY the same thing as each other, would they have all got the same results as each other? Explain why you said that.

Specific objectives

Probe 3 presented a task with a variable structure that was almost similar to the earlier probe, and this consisted of a single dependent continuous variable (percentage increase in mass) against two independent categoric variables (apples and potatoes); the variable design was deemed not too complicated for the PSTs.

The current probe, however, differed from the previous one in that the experimental data were presented, and therefore, variation was a given from the onset for all PSTs including those who did not see uncertainty as an intrinsic

property of measurements. The variation could be reasoned by claiming that the tabulated data were obtained from *twelve* individuals (collating data from different students is a common strategy in school-based investigations to circumvent limited time). In addition, the probe was set in a biological context, thus, since it involved living matter, the data might be more susceptible to uncertainties including the nature of the samples.

Probe 3 exemplifies the application of the triangulation strategy ("within-instrument") using a different context; it was designed to further investigate questions raised in previous probes (Probes 1 and 2): What caused variation in repeated readings? What other sources of random errors other than from humans could be the causes of variation in measurements?

Results and Discussions

Question 7:

After data reduction, the following categories were derived to show the distribution of the PSTs' reasoning:

(a) Human errors, for example, errors in preparation of sugar solutions, parallax errors in reading lengths(13);

(b) Natural differences in apples (uncontrolled variable), for example, the variety of apples, water or sugar content in different parts of the apple(22);

(c) Temperature of the solutions (uncontrolled variable)(1);

(d) Errors in the measuring instrument like the top-pan balances(2);

(e) Theoretical reasoning (2).

It is important to note that more than one type of error was sometimes given by the PSTs; less than half actually gave two, and a few, three causes of variation. As shown in the list, category (b), "natural differences in apples" was

the most common cause of variation followed by “human errors”. Contrary to the last probe, “environmental factors” (for e.g., room temperature) were conspicuously absent reflecting the PSTs might not routinely use “environmental factors” as a reason. The two PSTs in category (e) relied completely on their theoretical knowledge that variation must always be present.

Additionally, despite human errors being the major cause of variation in the previous probe, most individuals did not even mention it until Question 8 was posed, which indicated it was not as significant as before. Perhaps, the presence of another biological system (apples) had drawn the PSTs’ attention away from human errors. Still there were three PSTs who cited human errors as the *only* cause of variation in this probe, and among them, one PST was flagged before in Probe 2.

Question 8:

The imposed condition Question 8 that pupils did the experiment “exactly the same way” did not sway the PSTs as half (fourteen) maintained the readings would appear like before. This implied the PSTs in this category firmly believed that all causes of variation (including “natural differences of apples” and “human errors”) mentioned in Question 7 still persisted.

Ten PSTs in the second category claimed the readings would become closer because they expected the variation to reduce (but not totally removed):

> Perhaps the range will be closer; the results will be more similar. You cannot guarantee; they are not robots. (I1-2)

This implied the ten PSTs recognised that the twelve pupil-investigators were still an important source of errors despite them performing the experiment “exactly the same way”. The hypothetical imposed condition, however, might be translated to mean the pupil-investigators were deploying *good* measuring techniques that eventually led to a smaller degree of variation in the data.

The final category belonged to four PSTs who claimed the readings would become the same throughout ("zero" variation). Two of them had made similar claims before in the previous probes. The response from one such individual is shown below:

> I assumed the apple is the same but *it doesn't really matter*, and if the students did the experiment according to instructions, they will get the same results. (I1-15)

The four PSTs did not seem to understand that variation is caused by different sources of random errors (not just human errors); and, they probably believed that variation could be eliminated totally by the pupil-investigators performing the measurement in "exactly the same way" (perhaps, human errors were viewed as "mistakes" that could be corrected, but this needed to be investigated further), which unlike the second category of PSTs, might have been translated by the four PSTs to mean the pupil-investigators had "perfected" their measuring technique (and therefore, able to obtain a set of "perfect" data).

Review of Probe 3

The results from Question 7 indicated that the PSTs seemed to be quite knowledgeable about uncontrolled variables, and provided a range of possible factors that needed to be controlled, and their resulting errors; for instance, they mentioned the sugar or water contents of the apple chips, and the concentration of the sugar solution might be the causes of variation:

> Even though they say its equal size, I think there should [still] be some kind of error in the sugar solution… like it is too diluted or too concentrated. (I1-27)

A number of PSTs found Question 8 problematic because they could not easily accept the hypothetical condition imposed in the question. They also claimed the question was "tricky"; several PSTs required further explanation on what was meant by "exactly the same way" (particularly, whether the phrase included the errors committed by pupil-investigators).

The results of the probe also showed the PSTs expected living systems to be susceptible to uncertainty. This implied the PSTs might show a heightened sense of uncertainty when presented with biological-based investigative probes. The response below supports this observation:

> You can't get exactly the same reading for life experiments. (I1-10)

4.4.4 Probe 4

Probe 4 enabled me to explore the PST's understanding of variation in repeated readings in the context of having to establish the relationship between two variables in a Type 2 investigation in which collected data showed high uncertainties (thus, providing an authentic situation).

A scientist wanted to see how temperature affected osmosis in a potato. She puts equal size chips in a 1 mol/dm^3 sugar solution at different temperatures and measured the change in mass of the 'chips' on a top pan balance and recorded the results as a percentage of the original mass after 4 hours. She repeated all the readings 3 times.

Percentage of original mass			
Temperature (°C)	1	2	3
15	83.46	78.88	80.91
30	77.5	82.67	80.59
45	82.66	65.47	74.24
60	67.76	93.32	73.18

Question 9: What does the data show? Explain all your reasoning?

Specific Objectives

Probe 4 used the same context and topic ("osmosis") as the previous Probe 3 but it had a more complex variable structure: a single continuous DV (percentage of mass) against a single continuous IV (temperature). The one accompanying question was open-ended, which should allow a wide range of responses. The PSTs were requested to "think aloud" while they mull over the data. The objective of the probe was to investigate the question: How the PSTs would interpret a "messy" set of DV repeats against IV values that showed no trend or the data appeared "counterintuitive"? By being "messy" it meant the repeated DV values have a high degree of variation and are well spread out.

The high uncertainties in the DV data can be better understood by looking at columns 4 to 6[24] in Table 4.1.

Table 4.1 Data Table for Probe 4

Percentage of original mass				Standard Deviation (4)	Mean (of two closest values) (5)	Mean (of three values) (6)
Temperature (^{o}C)	1	2	3			
15	83.46	78.88	80.91	1.87	79.90	81.08
30	77.5	82.67	80.59	2.12	81.63	80.25
45	82.66	65.47	74.24	7.02	78.45	74.12
60	67.76	93.32	73.18	11.00	70.47	78.09

(a) As shown by the standard deviation values provided in (4), the variation in the DV was low at 15^{o} and 30^{o} but increased significantly for the next two IV values, which meant a high degree of variation existed.

(b) To check for trends, the majority of PSTs would likely begin by calculating the mean values using either two or all three DV values as indicated in columns (5) and (6). They might then try to generalise the relationship between the variables by comparing the columns for temperature against either columns (5) or (6). In both cases, they would not be able to find a distinctive pattern to draw any conclusion.

Results and Discussions

Question 9:

After coding analysis, the PSTs could be placed in two distinctive categories: those who examined and described the data *horizontally across* the table, they were called the “rowers”; and, those who did the same actions but vertically up or down the table, and they were labelled “columners”. To illustrate, I1-7 who was categorised a “columner” said:

> Now I’m *looking downwards* to see how when the temperature is increasing, how it affects the mass, I am looking for a pattern or a trend. (I1-7)

On the other hand, I1-5 who was categorised a “rower” responded:

[24] These columns were not given in the actual question but used here for discussions.

> *Looking across* the table, the values differ quite greatly for certain temperatures (I1-5).

After data reduction, ten "columners" and fifteen "rowers" were found; three PSTs did not belong to either category because they gave a mixed response such as the one below:

> I was looking down and I was looking at each temperature, looking for a relationship…I was looking down the columns 1, 2 and 3 individually but I am also looking across for 1, 2 and 3. I am trying to see if there's any relationship between the temperature and the [DV values], for example, 15 with 1, 15 with 2, and 15 with 3. (I1-9)

Several points could be noted from the "rowers" and "columners'" pattern of responses. Some "rowers" claimed the scientist did the DV measurements repeatedly in order to find similar readings across the table. This was consistent with the earlier findings in Probe 1. A good illustration was provided by I1-21's response:

> The results are not the same for all three although she repeated it three times…you repeat to ensure accuracy, so she [the scientist] did it three times and not once, [so that] she would be able to use them. (I1-21)

When I1-21 was asked to state her follow-up actions if she were the scientist, she replied:

> I would repeat it again to see how come it varies. You should get standard results for all three. You should get consistent results to conclude. (I1-21)

I1-21 exemplified a sub-category of "rowers" whose main objective was to find a consistent value for DV. Another "rower" response given by I1-6 showed similar motivation:

> I'm looking *across* the data… [The scientist] must have exactly the same temperature, and maybe put in [the potato chips] at the same time, and the same amount of everything in order to get the *same* data. (I1-6)

Like I1-21, the other "rowers" were generally unsatisfied with the three repeats, so they claimed they would perform more repeats if given the opportunity. A second sub-category of "rowers" had a different intention for the additional data; they aimed to have a better selection of close readings to calculate a mean

value, which would be used to represent the DV values responding to a particular IV interval:

> I'll do another, a fourth or fifth time so as to get an average value at the end, which is easier to see. (I1-1)

Finally, the PSTs who were flagged before for assuming that repeats should be invariant were all "rowers". They seemed to be very focused at checking to see whether the readings across the table were the same.

The "columners" had a different perspective on the scientist's purpose of performing repeats; they claimed she was looking for a trend between the percentage of the original mass and the temperature. The response below shows this:

> Now I'm looking downward to see when the temperature increases, how it affects the mass [of the potato chips], I am looking for a pattern or a trend. (I1-7)

Since the "columners" were also unable to find the relationship between the IV and DV, they were asked to imagine what follow-up actions they would take if they were the scientist. Two sub-categories of "columners" were found based on the responses. One generally suggested carrying out a whole set of DV measurements for the same range of temperatures:

> She put in the chips at 15…60, and after that, measure everything. Then after one hour she again does the same thing again. (I1-13)

Another would only replace "unsatisfactory" data points in the columns by repeating the measurements only for those points:

> I will not take into account the second one [reading] for 45 and 60^{o}C. I'll do again just to see whether they were inaccurate. (I1-4)

Review of Probe 4

During the interview, the presented data was effective in provoking a wide range of responses. The "thinking aloud" strategy allowed better "capturing" of the PSTs' ideas because their thinking was made "visible".

The rather set way of reading data reflects the PSTs' past learning experiences, which largely consisted of performing "cookbook" laboratory exercises (Kirschner & Meester, 1988; Clackson & Wright, 1992). Often, in such practical work, the priority was to develop routines (Toh, 1994) in order to arrive at the "right answers" (Fairbrother & Hackling, 1997). There was little emphasis on designing experiments as variable-type experiments primed to show trends were mostly given for practice (Kim, Tan, & Talaue, 2013). This explained why the "rowers" were very focused on getting a few DV readings so that they could use them to calculate a mean or select a value to show the expected trend.

An important question that could be derived from Probe 4 concerned how the "rowers" and "columners" would plan and perform their investigations. This could be the objective of another probe in subsequent studies.

4.5 Summary of Chapter

Table 4.2 below gives a summary of key findings from P1I1.

Table 4.2 Summary of P1I1 findings

Probe	Key Objectives	RA*	Key Findings
1	Purpose of repeated measurements	1	Consistent (or precise) data, mean value, for checking, or a routine practice to be followed. Only 7% aware that repeats "capture" data spread
1	Belief in true values	1	About 18%
2	Inherent variability of measurements	1	About 85% understood
2	Causes of variation in repeats	1	Human errors; environmental factors and other uncontrolled variables
2	Human errors as the *only* cause of uncertainty	1	About 25% (context–dependent)
2	Statistical concepts like standard error or standard deviation	1	No evidence shown
2	Precision can be improved using more reliable instrument	1	About 20% might not have understood
2	"Perfect" instrument/method of measurement can exist	1	About 25%
3	Uncontrolled variables a cause of uncertainty(variation)	1	Yes; more pronounced with living matter.
4	Processing "messy" tabulated data	1	"Rowers" and "Columners" (and possibly a third "Mixed" category)

*RA = Research Aim

CHAPTER 5

PHASE 1 QUESTIONNAIRE 1 STUDY

5.1 Chapter Overview

This chapter concerns Questionnaire 1, the second study in Phase 1 of this research, and shall be referred to as "P1Q1". The chapter begins by first stating the structure of the instrument plus its broad aims. This will then be followed by the main section that describes all the probes in three parts: the specification of objectives; the results and discussions of the analysis; and finally, a review of the probe. The chapter ends with a summary of P1Q1 findings. A copy of P1Q1 and a sample of a completed P1Q1 questionnaire can be found in Annex 3.2 and 3.9 respectively.

5.2 Structure of P1Q1 Questionnaire

The questionnaire consisted of three probes with the following focus:

Probe 1: procedural ideas on single measurements, repeated measurements, and variable-based investigations

Probe 2: uncertainty in a single measurement

Probe 3: uncertainty in data sets of repeated data (Type 1 investigation)

P1Q1 was completed by all fifty-five participants of the research who would be identified by codes assigned to them during the interview studies (P1I1 and P1I2).

5.3 Main aims of P1Q1

(a) The probes in P1Q1 were crafted based on leads provided by P1I1. This was done to triangulate some of the key findings surfaced in that study as well as to further explore the observed PSTs' understandings of uncertainty in *repeated* measurements. A new probe that investigated uncertainty in a *single measurement* by exploring the

choice of measuring instrument was also introduced. Thus, Research Aim 1 that was extensively explored in P1I1 remained the focus for P1Q1.

(b) The questionnaire developed for P1Q1 was intended to serve as the precursor for the final instrument. This allowed the design of some probes to be examined and tested in the light of developing "neutral" tasks and interpreting evidence based on "neutral ground" (see Section 2.3). In this respect, P1Q1 also targeted Research Aim 2 that spells out the intent of developing a questionnaire that describes the PSTs' understanding of uncertainty in measurements.

5.4 P1Q1 probes: objectives; results and discussions; review

5.4.1 Probe 1

Although the probe was titled "Repeats" and most of the fourteen statements looked at procedural concepts related to repeats, some also looked at procedural understandings related to measurements in general and variable-based investigation. There were two parts to the response for each statement: first, the PSTs were to indicate whether they agreed or disagreed with the statement; and second, they were asked to provide supporting remarks if they had any. The PSTs' responses were then checked against those provided by the experts[25] and the percentage agreement for a statement was used as an indicator of the overall level of understanding of the idea conveyed by the statement. As for the remarks, since they were given by the PSTs mainly to support their choice, they were used as further evidence to support the observed trends of thinking.

[25] The experts comprised my two supervisors and me.

Specific Objectives

Since the statements were mainly derived from the P1I1 findings, the responses would therefore meant to check earlier observations. Except for one (statement 13), a cluster of statements had been crafted to look at a particular idea/concept (and for the convenience of the reader, the clusters of statements will be shown together with the analysis of results in “Results and Discussions”). The ideas were fundamental to the understanding of uncertainty, and they included the inherent variability of measurements, the number of repeats, true value, and precision. A definitional statement on “fair test” was also posed to ascertain the PSTs’ knowledge of the essential condition imposed for variable-based investigation. The clustering of statements for analysis allowed comparison to be done across different statements. However, in the actual questionnaire, the statements were arranged randomly so that the PSTs would respond to individual statement rather than all of them together.

Results and Discussions

To get a profile of the PSTs as a group in terms of their understanding of repeats, a comparison of responses to statements 1 to 14 between the PSTs and the experts was done. The results are illustrated by Figure 5.1 and Table 5.1.

Figure 5.1 Number of PSTs against number of non-matching responses

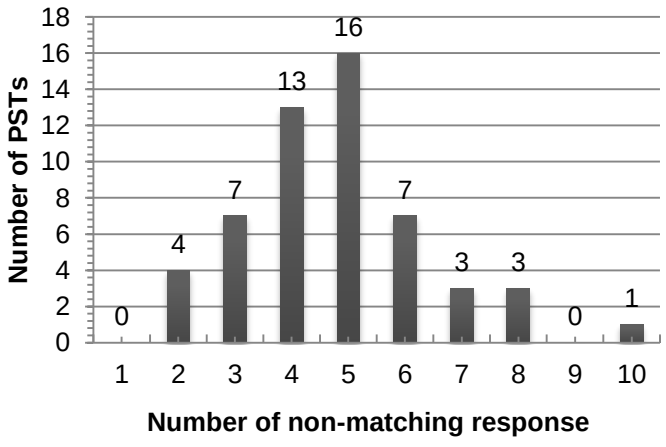


Table 5.1 The number of non-matching responses

Number of PSTs (N =55)	1	0	4	7	13	16	7	3	3	0	1
Number of non-matching response	0	1	2	3	4	5	6	7	8	9	10

The analysis showed that close to 80% of the PSTs had between three and six responses that did not match with those given by the experts (Mean =4.69, SD =1.75). Only one PST matched completely with the experts' responses. The next group of high-scoring individuals consisted of four PSTs who generally disagreed with the experts in statements 4 and 14 only. The ideas borne by statement 4 (consistent data should be the "right answer" for a measurement) and 14 (data sets can only be compared if they have the same number of data points) seemed to be well-entrenched since most PSTs gave similar response to both statements.

As mentioned earlier, Probe 1 was also analysed by looking at the PSTs' responses to each cluster of statements supporting a single idea/concept. The first of these was on the PSTs' understandings of the inherent variability of measurements (Table 5.2).

Table 5.2 Exploring understandings of the inherent variability of measurements

No.	Statements
7	Most measurements or readings when done several times would vary a bit no matter how careful you are.
8	It is never possible to repeat a measurement or reading in exactly the same way.

The results of the analysis are shown in Table 5.3.

Table 5.3 Results from exploring the inherent variability of measurement

Statement	Experts' response	% agreement[26]
7	Agree	94.5
8	Agree	87.3

The experts agreed with both statements 7 and 8. Based on the analysis of responses to statement 7 alone, we could see a higher percentage

[26] The percentages were rounded to 1 decimal place.

of PSTs (as compared to the previous P1I1 study) agreeing to the idea that repeats must *always* vary (only three PSTs thought otherwise). Looking at the cluster, the overall understanding that measurements would naturally vary whenever they were taken was quite high. On the other hand, a certain percentage of PSTs (12.7%) disagreed with statement 8 as they seemed to believe a reading could be repeated *exactly the same way*. In trying to understand their response, the PSTs' remarks were examined and they showed some PSTs might have "loosely" interpreted statement 8 to mean *steps* in taking measurements, which they assumed "could be done" (I1-54) or executed in exactly the same way.

The next cluster of statements that would be analysed concerned the number of repeats (Table 5.4).

Table 5.4 Exploring understandings of the number of repeats

No.	Statements
1	a. Two or three repeated measurements or readings are always enough. b. If you disagree, propose how many times would be enough: _____.
3	You should go on taking measurements or readings until you know what the range of a variable is.
5	*Ideally* you should take as many measurements or readings as you possibly can.
14	It is only fair to compare two similar experiments provided they have the same number of measurements or readings.

In P1I1, several PSTs showed understanding of the idea "the more measurements, the better"; so, the current study intended to explore this concept further (see Table 5.5).

Table 5.5 Results from exploring the number of repeats

Statement	Experts' response	% agreement
1	Disagree	76.4
3	Agree	76.4
5	Agree	87.3
14	Disagree	34.5

The high percentage of PSTs disagreeing with statement 1 indicates the group's non-inclination towards two or three repeats being sufficient for a

measurement. Of the 23.6% who agreed to the statement, nine PSTs gave a conditional remark basically stating that their responses were contingent on the "2 or 3 readings being consistent" (I2-34).

From those who disagreed, a small number remarked their reasons were based on the number of repeats being dependent on the variation in the data:

> Depending on how much each reading varied. (I1-54)
>
> Depends on how much errors were expected. (I2-32)

Nevertheless, a number of PSTs had also disagreed because they were thinking of another "fixed" number of repeats like five or ten. If the number of these individuals were lumped together with those who agreed to statement 1, then the actual number of PSTs thinking of a "fixed" number of repeats could actually be much higher than 23.6%. Most of these PSTs were probably falling back on their past laboratory experiences where the number of repeats were often pre-determined for them. Additionally, there were others who considered factors like the lack of time and the ease of planning instructions when they proposed a fixed number of repeats. In their study, Lee, Li, Goh, Chia & Chin (2002) have pointed out that such factors have strong influence on the PSTs' decisions in implementing inquiry-based activities in the local primary science curriculum.

Such practical considerations were also given in the remarks for statements 3 and 5 which meant the small percentages of PSTs who disagreed with the two statements could have likewise based their ideas on practical "realities on the ground". These PSTs placed more stress on operational considerations than on procedural understanding when they responded to the statements. One PST I2-11 might be emphasising such a notion when he wrote "the word *ideally* must be there" in his remarks to statement 5.

There was a small difference between the percentage agreements to statements 3 and 5. The lesser agreement to statement 3 could be the result of the PSTs not agreeing to the part of the statement that said: "…until you know what the range of a variable is", which was shown in P1I1 (and later in this current study) as not being the PSTs' purpose for repeating a measurement.

The responses to statement 14 showed a misconception existed with respect to the concept of a fair test [the concept appears to be challenging for teachers elsewhere as well; see Jarvis, Pell, and McKeon (2003)]. A number of PSTs assumed a valid comparison could only be made between two data sets if both had the same number of repeats.

In the previous P1I1 study, we saw about 18% of PSTs believed in the existence of true values. Such a belief could lead to other notions, for example, the existence of invariant repeats or a "perfect" instrument or measurement technique. The PSTs' concept of true values was further explored through three statements shown in Table 5.6; the results of analysing the statements are shown in Table 5.7.

Table 5.6 Exploring understandings of true values

No.	Statements
2	People who are good at doing experiments always get the *same* measurement or reading each time.
4	You know you have got the *right answer* when you are able to get the same measurement or reading twice or more.
12	We can perfect a measurement technique so that only one measurement will give a 'true' value.

Table 5.7 Results from exploring the concept of true values

Statement	Experts' response	% agreement
2	Disagree	89.1
4	Disagree	34.4
12	Disagree	69.1

The high percentage for statement 2 meant that the PSTs generally disagreed with the statement. This was supported by remarks such as the one given below:

> No matter how good they are, there are bound to be human errors. (I2-53)

But a small group of 10.9% believed otherwise and thought it was possible for investigators to achieve the same measurement if they were good at carrying out the task:

> They exhibited the *right* techniques when they did the experiments. (I2-55)

The concept of random human errors seemed to be poorly understood by such individuals.

Nevertheless, there were also PSTs who disagreed on the basis that uncontrolled variables could also affect measurements:

> It depended on the nature of experiment and what variables you could control. (I1-11)

On statement 4, the results reiterated a previous observation in that a high percentage of PSTs believed they got the “right” measurement when the same data was obtained twice or more. One might argue that the PSTs might have taken into consideration the limitations in the sensitivity of a measuring instrument or its resolution of scale but such notions were never demonstrated in the current study or in P1I1. Finally, the result for statement 12 concurred with those from Probe 2 in the P1I1 study that also showed about a quarter of the PSTs believed in the existence of a “perfect” measuring technique or instrument.

As described in Chapter 2, precision refers to the distance between the repeated measurements whereas accuracy is concerned with the gap between the measurement and the true value. But the fundamental difference was often overlooked by learners including the PSTs in P1I1 as they seemed to use the

term “accuracy” in place of “precision”. The cluster of statements that explored the PSTs’ understandings of the concept of precision is shown in Table 5.8 and the results of their analysis are shown in Table 5.9.

Table 5.8 Exploring understandings of precision

No.	Statements
6	If you get one measurement or reading that is very different from all the others you should ignore it.
9	Precision means the values obtained in repeated measurements or readings are clustered closely together.
10	The less an instrument’s precision, the more is its uncertainty.
11	A precise measurement may not necessarily be an accurate or ‘true’ measurement (and vice versa).

Table 5.9 Results from exploring the concept of precision

Statement	Experts’ response	% agreement
6	Agree	27.3
9	Agree	85.5
10	Agree	96.4
11	Agree	89.1

The experts opined that a datum identified as an anomaly should be discarded, and thereby agreeing with statement 6. But in Table 5.9, the percentage agreement for statement 6 shows a significantly low value. The result could be misleading as many PSTs seemed to agree with the statement but eventually disagreed because they felt the statement was incomplete and did not capture the notion embedded in the following responses:

> Find out what causes the anomaly. (I1-24)
>
> Don’t ignore but try to find out why it is so. (I1-1)

For the PSTs, an outlying datum should not just be ignored; rather, it must be further investigated to determine what really caused the anomaly.

The results for statements 9, 10 and 11 collectively show the PSTs understood the meaning of “precision” and that the term is theoretically different from “accuracy”. The results confounded the P1I1 findings that generally showed the PSTs were not able to distinguish between precision and accuracy.

Perhaps, the PSTs were only able to define the terms and routinely recall what they represented (as demonstrated in P1Q1) but were unable to *apply* the concepts (as demanded in P1I1).

Statement 13 was included in Probe 1 to see if the PSTs understood the design of a variable-based investigation (that most probes in the study are based on), and that it involved establishing the relationship between two variables (which meant all other variables needed to be controlled) (see Table 5.10).

Table 5.10 Exploring understanding of a fair test

No.	Statement
13	A fair test is one in which only the independent variable (whose values are changed constantly) has been allowed to affect the dependent variable.

The result of analysing statement 13 is shown in Table 5.11.

Table 5.11 Results from exploring the concept of fair test

Statement	Experts' response	% agreement
13	Agree	92.7

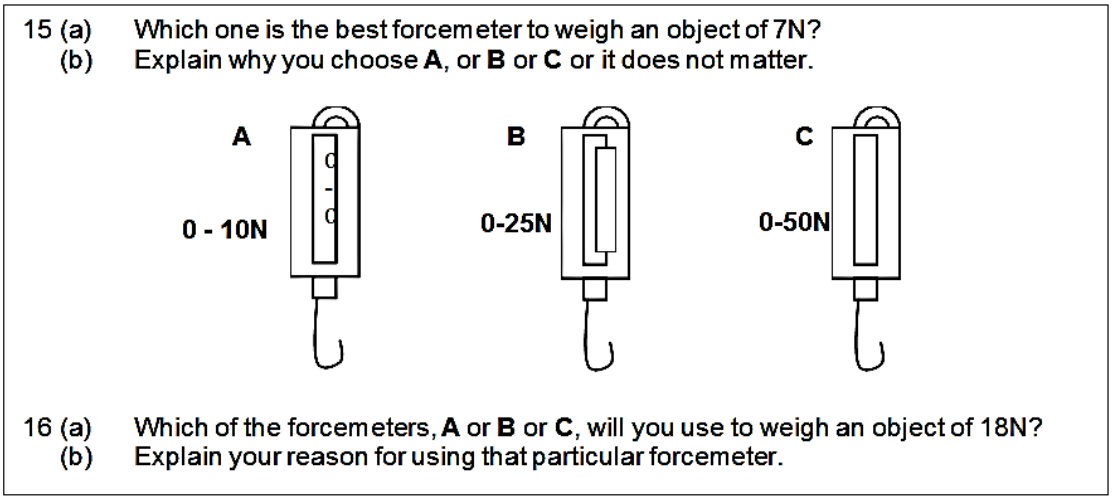
The high percentage agreement for statement 13 shows the majority of PSTs understood the definition of a fair test (which could lead to recognising the effects of uncontrolled variables on measurements). However, the definitional knowledge of a "fair test" did not seem to be correctly applied in the light of results to statement 14 (see Table 5.5). There were a number of PSTs who erroneously extended the idea of "fair" to include having to consider the number of repeats when comparing data sets. This will also be evident in other probes later. The concept of "fair test" like "accuracy" and "precision" might be ideas the PSTs had routinely learned without any real understanding; they may well be examples of routine knowledge that had become "inert" (Perkins, 2006).

Review of Probe 1

Although many of the statements had been used in other studies, and the questionnaire was piloted as well as vetted by experts including those from NIE, problems of misinterpretations by the PSTs still occurred. A good example was seen in the responses to statement 6, which seemed to be ambiguous and resulted in many PSTs suggesting an anomalous datum should not be ignored but further investigated to determine the cause of it being an outlier. Others, for instance, statements 3 and 8, might have similar problems in being ambiguous albeit for a relatively smaller group of PSTs. Despite reservations, the evidence from Probe 1 generally supported the results and interpretations of the PSTs' understandings of repeats from the P1I1 study.

5.4.2 Probe 2

Probe 2 investigated the PSTs' understanding of uncertainty in measuring a single datum. The concerns were largely on the uncertainties in a forcemeter and whether the instrument could give an accurate measurement of weight.



Specific objectives

The objective of the second probe entitled "Instruments" was to find out how the PSTs would decide on the best instrument to measure a certain

quantity [weight in Newtons (N)]. As described in Section 2.2.4 in Chapter 2, the choice of a measuring instrument like a forcemeter which has a fixed scale may be based on two interrelated concepts, “full-scale deflection (FSD)” and the “resolution of scale”. In a practical situation, the PSTs might also be checking for “zero errors” (Section 2.2.4) and “instrumental reliability” (perhaps, by carrying out “preliminary trials” with an object of known weight) (Section 2.2.5), but in order not to complicate the issues that are dealt in the probe, we shall assume these were already done. Another assumption would be the investigator knew how to use the given equipment and read their measurements so that these would not be held as factors for choosing an instrument[27].

Looking at Question 15(a), the best forcemeter to measure 7N should therefore be forcemeter **A** based on either the concept of FSD or the resolution of scale. For Question 16(a), it would be forcemeter **B**. The same concepts were deliberately tested twice to substantiate the evidence and bolster the findings (in the spirit of triangulation).

Results and Discussions

The results to Questions 15 and 16 are shown in Table 5.12. It is important to note that fifty-four PSTs (exception of one) responded with some form of explanation about why they picked the option, which were then used to determine the prevailing concepts used by the PSTs.

[27] These assumptions would apply to all probes in this research that dealt with the choice of instruments as well.

Table 5.12 Results of Probe 2 "Instrument" (N=55)

Q	Code description	Number	Total
15	**A**, resolution of scale	23	30(54.5%)
	A, FSD	7	
	A, both	0	
	A, obscure reasons	14	25.5%
	B or **C**, with reason	10	11(20.0%)
	No response	1	
16	**B**, resolution of scale	8	19(34.6%)
	B, FSD	10	
	B, both	1	
	B, obscure reasons	29	52.7%
	C, with reason	6	7(12.7%)
	No response	1	

The results show the PSTs often used the resolution of scale to choose their forcemeters. This is illustrated below for question 15:

> **A**, the markings on the forcemeter should allow for a more accurate reading, for example 7.45N, that may not be reflected on larger scales [of other forcemeters]. (I1-11)

Likewise, for Question 16:

> **B**, it gives you a more accurate measurement as the markings for a smaller range [scale] would be more. (I1-22)

In comparison, the concept of FSD was less applied. The responses below illustrate how two PSTs used the "FSD" to answer Question 15:

> **A**, the range is smallest thus 7N is closest to it. (I1-21)

For Question 16:

> **B**, it is closest to the weight of the object. Forcemeter **A** has a range that is too narrow. **C** has a range that is too wide. **B** is most ideal. (I1-51)

Based on the analysis in Table 5.12, many PSTs (about 80%) were able to pick the right choices, but supported with obscure reasons, particularly Question 16. This implies the PSTs might have tacit understanding in choosing the forcemeter based on the idea of "fitness of purpose" but were unable to clearly articulate the concepts of FSD or resolution of scale (which therefore gave rise to many obscure reasons). The problem seemed to be compounded by the lack of knowledge of technical terms to describe the concepts.

For those PSTs who picked the incorrect choices for both Questions 15 and 16, they often could not see the difference between the three forcemeters in terms of their scales. This can be seen in the response to Question 15:

> It does not matter. They are all forcemeters and they measure an object in Newtons. (I1-7)

They were also those who preferred using forcemeters with wider scales as they claimed the given measurement would fall within or near the centre of the scale, and as such, would allow the investigator to read the measurement conveniently:

> 7 Newtons will fall roughly around the mid-range [of forcemeter **B**] whereas for **A** and **C**, the reading will fall at the upper range and the lower range respectively, hence, reducing accuracy. (I2-37)

Review of Probe 2

A large part of the obscure responses in Question 16 came from individuals who could have simply repeated the same reason they gave earlier for Question 15; instead, they provided one that was lacking in details. Perhaps, the PSTs assumed question 16 was merely a repeat question that did not deserve a full explanation. Another problem was that the PSTs who relied on the resolution of scale to choose the best forcemeter in Question 16 might not be able to distinguish forcemeters **B** (0-25N) and **C** (0-50N) (see results for Question 16 in Table 5.12). This could have also caused several PSTs to pick both forcemeters **B** and **C** to measure 18N accurately.

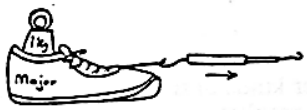
Finally, although the majority of PSTs was able to pick the best instrument to measure the given weights accurately, and seemed to understand the concept “fitness of purpose”, it remains to be seen whether the PSTs were able to apply similar conceptual understandings to other scale-based and digital instruments.

5.4.3 Probe 3

Probe 3 entitled “The Sole Test” used a simple physics investigation based on the concept of force and consisted of measuring a continuous variable (the pulling force) against a categoric variable (different ground surfaces).

The Sole Test

Kumar, Ahmed and Lee did some work about how different surfaces affect the ‘slippiness’ of a running shoe by putting a 1 kilogram mass in the shoe and finding the amount of pull needed to drag the shoe along. They tested each surface twice.



Here are their results:

Type of surface	*Pull force(Newtons)*	
	1st trial	*2nd trial*
Soil on the school’s playground	10	15
Grass on the school field	14	13
Carpet in the school’s library	8	9
Cement floor in the school canteen	5	7

The accompanying questions will be given separately

Two important characteristics were planted as distractors in the data: (a) “15” in the soil readings was comparatively higher than the first reading, and those from the grass; (b) the data from the soil and grass overlapped, and could cause difficulty in distinguishing which particular surface would result in the higher pulling force.

Specific objectives

The third probe looked at how the PSTs would apply their understandings of repeated measurements in the three essential parts of an investigation: data collection, data processing, and data interpretation. Table 5.13 below gives the specific objective of each accompanying question found in the probe. As seen in Table 5.13, Probe 3 basically looked at similar concepts explored in P1I1 that had used a different investigative task (the bouncing ball investigation). Thus, the results of the current probe would serve to triangulate and refine earlier P1I1 findings.

Table 5.13 Specific objectives to the accompanying questions of Probe 3

Q	Accompanying questions	Specific Objectives
17	Why did they test each surface twice? (*Data collection*)	What was the purpose of repeating the measurements?
18	They thought they had done everything the same but they did not get the same results. Suggest why. (*Data interpretation*)	What caused the variation in the repeated readings?
19	Their teacher asked them which surface needed the most force to pull the shoe along. Tick (√) the one you most agree with: □ Kumar said it was the grass □ Ahmed said he couldn't tell □ Lee said it was the playground Why did you choose that one? (*Data processing and interpretation*)	In drawing a conclusion, what procedural ideas were used in comparing and selecting overlapping data?

Results and Discussions

Question 17:

Fifty PSTs gave only one supporting idea while five gave two. The categories that emerged from the responses were as follows:

(a) For getting consistent readings(18);

(b) To check previous reading(17);

(c) To get a mean(16);

(d) To check for human errors(5);

(e) To find the spread(3).

To illustrate the PSTs' answers, an example of a response for each category of reasoning ideas is given below in Table 5.14.

Table 5.14 Examples of response for different categories in Question 17

No.	Reasoning idea	Examples of PSTs' response
(a)	For getting consistent readings	Probably to investigate whether the same result would be obtained which might further enhance the accuracy of their investigation if the results obtained were constant. (I1-9)
(b)	To check previous reading	To counter check their first trial for more accurate results. (I1-18)
(c)	To get a mean	To obtain multiple readings so that an average can be obtained. (I1-1)
(d)	To check for human errors	To minimise any possible measurement error due to human activity. (I2-52)
(e)	To find the spread	More than one experiment [datum] provides you with more range of the result. (I2-36)

The results for this analysis were consistent with those from P1I1. The observations made for each category were quite similar; for instance, in category (a), many PSTs gave “accuracy” as a mechanical response without elaborating what it means; nevertheless, upon closer examination, their responses seemed to suggest “precision” or the closeness between repeats. The PSTs generally related the idea that with more precise data, the results would become more reliable [see response (a) in Table 5.14]. Category (b) response could be linked to the purpose of verifying the previous result. It may well be that some PSTs in both categories (a) and (b) could be seeking a true value; this, however, could not be established because the probe was not designed to explore such understanding.

A high number of PSTs in category (c) suggested “getting a mean” as the reason why the measurement was repeated twice but how much of the responses were rote remained inconclusive. Category (d) “to check for human error” underscored the notion held by many PSTs that the measurements in the investigation were susceptible to human errors as they were taken presumably by three young and inexperienced primary students. Finally, we can see in category (e) just as we did in the P1I1 study that only a small number of PSTs understood the real purpose why measurements are to be repeated.

Question 18:

From the coding analysis, only four of the fifty-five PSTs explained with two reasons whilst the rest gave only one. Their ideas fell mainly into two large categories; a third “miscellaneous” category existed with two ideas, one was on theoretical perspective, and the other, on instrumental error. All other causes of variation could be grouped into two categories:

(a) Human errors(30);

(b) Uncontrollable factors (27).

Many of the coded responses in both categories gave a sense that the errors were randomly affecting the measurements of force.

The fact that human errors were frequently cited supported earlier findings from P1I1 that human errors were often viewed as the main (sometimes, the *sole*) cause of variation. However, the results should be interpreted cautiously since the context of the measurement seemed to lend support to human errors being the major cause of variation (as mentioned earlier, the task involved not one but three different young and inexperienced primary student-investigators). Such a notion was implicit in several responses including the one below:

> The way they might have pulled the shoe could be different (e.g. the angle), maybe different people applied different force. (I2-32)

There was only a slight difference between the numbers of coded responses in (a) and (b) reflecting that uncontrolled variable was equally important. Ideas given in category (b) "uncontrolled variables" included: the surfaces might not be level and contained bumps; altering environmental conditions such as "wind resistance against the [pulling of the] shoe"; and, changes in the soles of the shoes as a result of wear and tear due to multiple measurements.

Question 19:

The responses to Question 19 were first categorised according to the three offered choices shown in Table 5.15. The number of PSTs who opted for each choice and examples of response given to justify their choices are also shown in Table 5.15.

Table 5.15 Results of Question 19 (N= 55)

Preferred choice (frequency)	Examples of response
Grass (34)	The average reading for the pull force on the grass was the highest at 13.5N. (I1-1) The readings between first and second are closest as compared rest, it has the closest large pull force measurement. (I1-21)
Playground (6)	Soil has the most amount of friction due to small sand particles. (I2-38)
Couldn't tell (15)	The range of the results for the grass and the playground are not defined from each other and the range overlaps (fall under common range). (I2-29)

As the range of responses supporting each choice seemed to contain similar ideas; on further analysis, a pattern of ideas can be found for each one (see Table 5.16).

Table 5.16 Categories of responses to Probe 3 Question 19 (N=55)

Choices (frequency)	Categories of reasoning ideas (frequency)*
Grass (34)	• The mean calculation showed highest for grass(21) • The total pull force(14 + 13) was highest for grass(4) • The first and second trial were closest(8)
Playground (6)	• Prior understanding (for e.g., the soil should cause the most friction and therefore the pull force must be the highest)(6)
Couldn't tell (15)	• Overlapping readings in grass and playground(8) • Insufficient trials(4)

*The shortfall in the total frequency were due to "uncodeable" ideas

The best choice for the question should be "couldn't tell" since there were too few readings and the data from both grass and playground overlapped. Nonetheless, only 27.3% (15) opted for this choice.

Six PSTs chose the playground because they gave priority to their own constructed knowledge about surfaces (for e.g., soil should give the highest frictional force because it was the "roughest") over the possible inferences from the presented data. This concurred with Masnick and Morris (2008) who claimed that learners tended to hold on to their own theory and might discount data that did not match with their prior understanding. Additionally, the readings

for the playground included the highest pulling force of 15N and this might have strongly distracted a number of PSTs.

Close to 40% of the PSTs chose grass on the basis that its mean value of “13.5” was the highest although this was not significantly higher than the mean value for the data from the playground (“12.5”). It was not surprising that those who used this idea earlier claimed in Question 17 that repeats were meant to provide a mean value. Additionally, eight PSTs in the same category saw the smaller variation in the data for grass compared to playground as a justification for their choice. They assumed the data were more reliable given the values appeared to be closer. Some members of this group also rejected “soil” as an answer because they believed the 15N pulling force could be a fluke. Interestingly, four PSTs who chose grass simply added the two readings to derive a total value as a measure of the highest pulling force. Their rather odd way of determining the highest pulling force could have been acquired from their past learning experiences.

Review of Probe 3

In my post-survey reflection, there was misgiving about the use of only two data points in Probe 3 as this could mislead the PSTs into thinking that two readings were sufficient for an investigation. Thus, if the probe were to be used again in the final questionnaire, it must be modified to remove this misrepresentation. Still, the probe was useful in revealing the PSTs’ reasons for carrying out repeats and it reaffirmed results largely from Probes 1 and 2 in P1I1.

Question 18 sought to explore the PSTs’ causes of variation in a different investigation task and to see if the PSTs would pin it down to only human errors. Human errors were indeed being cited as a major cause of

variation; however, this could have been influenced by the task context. The critical learning point here is that the context of the probe could be a strong influence on the participants' use of an idea such as "human errors" when responding to a probe.

Finally, the results of Question 19 indicated only about 15(or 27%) were uncertain about the measurements and their doubts were largely over the number of repeats, variation of readings, and the fact that the data sets overlapped. A few PSTs in this category went further and suggested that more repeats should be necessary to draw a conclusion. This implies some understood the idea "the more measurements, the better" but whether this knowledge was based on statistical concepts could not be clearly ascertained from the response.

5.5 Summary of the Chapter

The results of the probes in P1Q1 mostly reiterated earlier findings from P1I1 (see Table 4.2 in Chapter 4). The key P1Q1 findings are summarised in Table 5.17 below.

Table 5.17 Summary of key P1Q1 findings

Probe	Specific objective	RA*	P1Q1 findings	Compared to P1I1 findings
1	Inherent variability of measurements	1	Between 85% and 95%	Agreeable
1	Belief in true values	1	Between 11 and 31%;	Agreeable
1	Number of repeats dependent on data spread	1	Between 76 and 87%; some believed data sets with only equal number of data can be compared	New item
1	Concept of precision	1	Between 85 and 96%. Anomalous reading must not be ignored but investigated for the reason of its deviation	New item
1	Concept of fair test	1	About 93%	New item
2	Selection of best measuring instrument	1	Most PSTs selected based on “fitness of purpose”. The PSTs mainly used the resolution of the scale over full-scale deflection in making their final choices.	New item
3	Purpose of repeats	1	Mostly for precise data; to check previous reading; and to get a mean	Agreeable
3	Causes of variation in repeats	1	Human errors and uncontrolled variables (context-dependent)	Agreeable
3	Procedural ideas used in comparing and selecting overlapping data	1	Mean value; variation in data; substantive knowledge (in order)	New item

*RA = Research Aim

CHAPTER 6

PHASE 1 INTERVIEW 2 STUDY

6.1 Chapter Overview

This chapter reports Interview 2 ("P1I2"), the third and final study in P1. The chapter begins by describing the interview protocol and the main aims of P1I2. The main section comes next looking at one probe at a time: its specific objectives, results and discussions, and finally, a review. The chapter ends with a short conclusion of P1. A copy of the interview protocol and a transcript from P1I2 can be found in Annex 3.3 and 3.8 respectively.

6.2 Structure of P1I2 Interview protocol

The instrument could be divided into two parts: the first consisted of six short probes that were either general or task-specific; and the second comprised of three longer probes, each looking at several procedural concepts. The focuses of the probes are shown below.

Probe 1 to 5: procedural concepts related to measurements

Probe 6: uncertainty in a single measurement

Probe 7: uncertainty in a single set of repeated data (Type 1 investigation)

Probe 8 and 9: uncertainty in different sets of repeated data (Type 2 investigation)

P1I2 was carried out with the *remaining* 27 PSTs. Due to some technical errors, two recordings were discarded; therefore, the P1I2 findings were based on 25 PSTs.

6.3 Main aims of P1I2

The probes in P1I2 were generally intended to triangulate findings from the two earlier studies. Several probes from P1I1 were retained in the current study to see how another group of PSTs would respond to them. However,

following the review of P1I1, some modifications had been made to the probes so that the PSTs in the current study would be able to interpret and respond accordingly. Therefore, the differences in the results of this study might not only be due to a different group of PSTs responding, but also the improvements made to the probes. New probes had also been designed to present other ways of asking, and to see how certain patterns of understanding (for instance, "columners" and "rowers"), represented the PSTs' ways of thinking.

6.4 P1I2 probes: objectives; results and discussions; review

6.4.1 Probes 1 to 4

For the convenience to the reader, the descriptions of Probes 1 to 4 would be given later in the "Results and Discussions" section.

Specific Objectives

Probes 1 and 2 questioned the PSTs' operational definitions of "accuracy" and "precision", two fundamental concepts that underpinned uncertainty in measurements. The P1I1 study (see Section 4.4.1) showed the PSTs' had a rather "loose" meaning of "accuracy"; it was often used inappropriately to replace "precision" in repeated measurements. However, in P1Q1 (see Section 5.4.1), a high percentage of PSTs chose the right statements that described "precision" indicating a high level of understanding. The findings so far seemed to imply the PSTs might have tacit knowledge of "precision" but could not articulate their understanding of the concept during the P1I1 interview. To prove this, first the PSTs would be directly questioned on their knowledge of both concepts in Probes 1 and 2, followed by Probe 3, which would require the PSTs to apply their understandings of the concepts to a familiar scenario. Probe 4 would then explore the PSTs' knowledge of the four situations where the two concepts could be interrelated (see Section 2.2.2).

Understanding, for instance, that precise measurements might not necessarily mean accurate readings would be critically important in evaluating the quality of data, and deciding the procedural actions that needed to be taken.

Results and Discussions

Probes 1 and 2:

- A reading is described as being an "accurate" value. Can you explain what do you understand by this statement?
- What do you understand by the term "precision" as applied to repeated measurements?

For Probe 1, most PSTs gave a single meaning of "accurate value" but three gave two. The coding analysis reduced the data to six categories shown below:

(a) Close to the established/literature value(7);

(b) Least affected by errors(3);

(c) Obtained by the correct procedure(5);

(d) Many decimal places(5);

(e) Mean value of repeats(7);

(f) True value of a measurement (2).

Category (a) showed only seven PSTs described their understanding of accurate value giving the accepted definition (see Section 2.2.2). Others from (b) to (e) had an incomplete understanding of the term, or even a misconception as shown in the response from (f) below:

> Accurate value is the *real* value, which I imagine is the true value out there that I can compare if there's any discrepancy. (I2-49)

For Probe 2 on "precision", one PST claimed from the outset that she did not know its conceptual meaning, and another declared it had the same meaning as "accuracy". The responses from three PSTs were too vague to be coded and categorised.

The remaining PSTs gave mostly a single meaning of “precision”, but two provided two different meanings. After iteratively going through the coded responses, six categories emerged:

(a) Closeness between readings(3);

(b) Most reliable measurement procedure(9);

(c) Least affected by errors(2);

(d) Many decimal places(2);

(e) Mean value of repeats(5);

(f) True value of a measurement (1).

Category (a) showed only three PSTs gave an acceptable definition of “precision” (see Section 2.2.2). Overall, in both Probes 1 and 2, only one PST gave an acceptable description of both concepts. Nevertheless, the categories that emerged for “precision” looked similar to those obtained for “accuracy” but a check done on individual transcripts did not reveal a PST who had given similar response to both probes, which basically implied the PSTs knew the two concepts were different.

Question 3:

A scientist reported that she obtained several melting point values for an organic substance and all are close to the value reported in the Data Booklet published by the Department of Science at the University. Describe the set of melting point values in terms of their accuracy and precision?

Probe 3 was based on the common knowledge that we could use melting point to determine the purity of a chemical substance. The intended answers would be the melting points were accurate (because they were close to the value reported in the Data Booklet) and precise (as the values were near each other). PSTs with difficulties differentiating the two concepts might feel challenged as they were required to explain both in the same probe. The

responses from seventeen PSTs described the melting points as accurate and gave acceptable explanations to support their answers; one is given below:

> If it is within the range given in the data booklet, it would be accurate. If they repeated the experiments, I would say it's accurate if they're close to the data in the booklet. (I2-32)

The remaining eight PSTs differed by claiming the readings were inaccurate as most expected the melting points to be exactly the value reported in the Data Booklet:

> If it's accurate, it should be the same as the one reported in the data booklet. (I2-30)

On precision, the results showed only one PST correctly described and explained the set of readings as precise. The other PSTs were generally vague, and some eventually admitted they did not understand the concept. Nine responded that "precision" was concerned with adopting good measuring procedure, and one claimed precision had to do with the number of decimal points. Four PSTs felt the melting points were not precise because the readings "were not the same (I2-34)" and they "should be the literature value (I2-38)".

Question 4:

Refer to statements I to IV below. Based on your understanding of 'accuracy' and 'precision', which describes the possible relationship between the two terms for a set of repeated readings? You may choose any combination as your answer.

I. Repeated readings can be both accurate and precise.
II. Repeated readings can be accurate but imprecise.
III. Repeated readings can be precise but inaccurate.
IV. Repeated readings can be both inaccurate and imprecise.

The intended answer was all four were possible but only six PSTs gave such a response. For the rest, 80% picked options I and IV only, and the remaining 20% picked three, usually I, IV plus either option II or III. The results showed that most PSTs did not have enough understanding of the two concepts to recognise their interrelationships. Further, the pattern of selecting the options

underscored the thinking that accuracy and precision should always be in agreement, thus an oppose situation like option II ("precise but inaccurate"), would unlikely arise.

Review of Probes 1 to 4

Although the generalised probes were able to draw the PSTs' focus to the concepts in question, most PSTs struggled to verbally articulate their understandings. Many chose to cite examples to draw the differences between "accuracy" and "precision" but their explanations often became vague and convoluted. Several PSTs actually ended up being even more confused by their own explanations. Nonetheless, it demonstrated how the PSTs might try to make sense of a probe with a generalised context by imagining a measurement situation.

The results for Probe 3 that presented a familiar task-specific context showed most PSTs seemed to have tacit understanding of only accuracy but not precision. The overall findings from the four probes and from previous P1 studies showed the PSTs might not fully understand the meanings of certain concepts related to uncertainty in measurements. Although "accuracy", "precision", and to some extent, "fair test" (see Section 5.4.1), were investigated, similar problems might exist in the understandings of other critical concepts like errors, variation, etc.

6.4.2 Probes 5 and 6

Specific Objectives

In P1I1 (Section 4.4.3), we saw the PSTs recognising "uncontrolled variables" aside from human errors as an important cause of uncertainty. Likewise, Probe 5 would look at how important the errors from measuring

instrument relative to other sources of errors particularly human errors (see Sections 4.4.2 and 6.4.3) in contributing towards uncertainty.

The understanding of uncertainties in single measurement had been investigated through the choice of instrument (see Probe 3 in P1Q1; Section 5.4.2). Probe 6 would investigate further issues raised by the probe; in particular, the difficulty in which many PSTs seemed to have in expressing their reasons for choosing an instrument. Given a different but familiar context of laboratory beakers, perhaps the PSTs might be able to demonstrate clearer understanding of the concept/s that led to their choice of instrument.

Results and Discussions

Probe 5:

> If scientists use the same instrument several times to take repeated measurements of the same variable, would they get *exactly* the same reading? If the answer is no, explain why?

The intended answer would be "no" due to all sources of random errors. In P1I1, the phrase "use the same instrument several times" was not used so the PSTs might think the errors came only from the scientists. With its inclusion, perhaps the PSTs might consider the contribution from instrumental errors.

However, the coding analysis showed not all PSTs claimed the measurements would vary; four claimed the repeats would all have the same value, and the response below came from one of those PSTs:

> Same variable, same instrument, hmmm…there's a high chance the readings will be the same. (I2-38)

From the remaining twenty-one PSTs who said "no", four attributed variation to two or even three different errors (see Table 6.1).

Table 6.1 Numbers and types of errors

Causes of variation	Human errors	Instrumental errors	Environmental factors
Frequency	11	8	5

Although human errors had often been highlighted as a source of errors in the probe, the results indicated the PSTs had considered other errors too.

Probe 6:

When fully melted, 80g of ice will be about 80cm^3 of water. Of what capacity beaker would you use for the ice: 500cm^3, 250cm^3, 100cm^3, or it does not matter?

The intended answer to Probe 6 based on "percentage errors" (see section 2.2.4) should be the 100cm^3 beaker since 80g of ice once completely melted would form 80cm^3 of water, which would appear near the top of the scale. Besides, 80cm^3 of water should be accurately measured by the 100cm^3 beaker as it had the smallest divisions of scale among the three beakers.

Twelve PSTs chose the 100cm^3 beaker and gave appropriate reasons like the beaker would have the greatest division of scale (8), the volume was near the top of the scale (3), and a combination of both (1). However, six PSTs chose the right beaker but could not give a clear explanation of their supporting concepts, thus highlighting a similar problem observed earlier in P1Q1 (see Section 5.4.2). Of those who chose the other beakers, four claimed it did not matter as long as the beaker could contain 80cm^3 of water, three chose *one* of the larger beakers because they felt "safe" using a beaker whose capacity exceeded 80cm^3 by a large margin.

Review of Probes 5 and 6

Despite the modification made in Probe 5, many PSTs still regarded "human errors" as an important cause of variation with respect to other sources of errors. Probe 5 also indicated the wordings in probes needed to be carefully crafted in order to get a "fuller" response especially if the response was intended to contain several factors, and not the main one only.

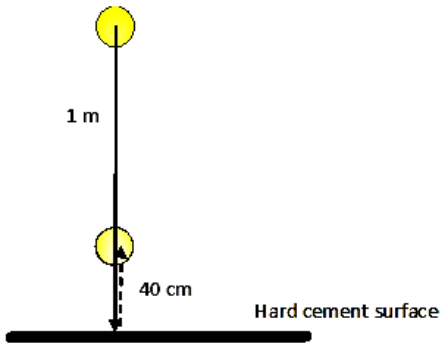
Probe 6 showed the PSTs might have difficulty explaining their tacit understandings of the concept of "fitness of purpose". Looking retrospectively

at the probe, the seven PSTs who chose a larger capacity beaker might be misled into thinking they needed only to suggest a beaker to be used as a container rather than to measure the volume accurately.

6.4.3 Probe 7: The "Bouncing ball" probe

Probe 7 was first used in P1I1; the current probe in P1I2 contained some modifications including new questions to investigate several emerging issues.

Imagine you simply drop a rubber ball from the height of one metre and then measure its rebound height against the metre rule. It bounced back to 40 cm the first time.



Imagine you bounced it 9 *more* times from the same height and each time taking note of the first bound height.

What would the 9 other first bound heights look like? Fill in imaginary results into the table below.

Bounce number	1	2	3	4	5	6	7	8	9	10
Bounce height (cm)	40									

Part of question 7 is shown; all other accompanying questions will be given later

Specific Objectives

The current probe retained several objectives described in P1I1 (see Section 4.4.2) plus a few new ones. In the first Question 7[28], the PSTs were asked to fill a table with their predictions of the next nine rebound heights. We could therefore see if the PSTs expected their repeats to vary, and if so, what would be the cause of it. The findings would be compared with those from

[28] For easier reference, the question was numbered in running order starting from the first probe in the interview protocol.

Section 4.4.2 to see whether human errors remained the main causes of variation. The PSTs would then be asked if they would continue to take readings after the tenth repeat and to explain why. This would tell us whether the PSTs based their decisions on the degree of variation or practical considerations (for e.g., time; see Section 4.4.2), or on a fixed number of repeats (Section 5.4.1).

The next Question 8 required the PSTs to use the data from Question 7 to report the "bounciness" of the ball. If the rebound heights were near each other, the most appropriate way should be to calculate the mean and the standard deviation. Most PSTs, however, might suggest only the mean value, which would be acceptable given this being the normal practice in primary science. Based on Section 4.4.1, there could also be PSTs who would select a consistent value (the mode) either because they understood it as representing a reliable datum or because they thought it represented a true value.

In Question 9, two data sets (containing overlapping values) with the same mean value (or "bounciness") derived from different number of repeats were being compared. The question would explore which characteristics of the two data sets (the number of repeats, consistency of values, smaller variations, and the presence anomalous result) would influence the PSTs' decision in choosing the more reliable data set.

Finally, Question 10 was meant to see whether the percentage of PSTs who believed in the possibility of a "perfect" instrument could actually be higher than found (25%; see Section 4.4.2). The percentage obtained in P1I1 might be affected by PSTs who appeared to have overemphasised the errors from the (imagined) human operator of such an instrument. To prevent the latter, the phrase "self-automated" would be introduced into Question 10. In addition, the

question also intended to explore whether the PSTs understood the introduction of a reliable measuring instrument (a digital camera) consequentially meant less number of repeats would be required.

Results and Discussions

Question 7:

> Why is the data varied? Ask if they would continue to take more data after bounce number 10? When will they stop, and why stop at that bounce number?

As illustrate by Figure 6.1 below, twenty-two PSTs gave almost similar response.

Figure 6.1 A typical response to Question 7

Bounce number	1	2	3	4	5	6	7	8	9	10
Bounce height (cm)	40	40	39	38	40	36	38	39	39	40

The nine data that were entered were normally near each other; a few PSTs also entered one or two outliers perhaps just to show the possibility of obtaining such data. There were only three PSTs who claimed the rebound heights should remain the same value throughout (see Figure 6.2).

Figure 6.2 Response from I2-53 showing no variation in the repeated rebound heights

Bounce number	1	2	3	4	5	6	7	8	9	10
Bounce height (cm)	40	40	40	40	40	40	40	40	40	40

The response below summed up why the three PSTs gave such a response:

> It's from the same height, the surface is also the same, so nothing changes, and the ball is also the same. Thus, it will bounce back to the same height. (I2-50)

There were also PSTs who gave short sequences of similar values, for example, I2-49 had "40" in the final few measurements, and I2-38 had "36" and "34" appearing recurrently in clusters. Apparently, all these were meant to show the data were reliable, and the investigation could stop.

On the causes of variation, there were about equal number of PSTs who cited one or two causes of variation. After multiple readings of the coded responses from the twenty-two PSTs, two large categories emerged as shown in Table 6.2:

Table 6.2 Causes of variation in height measurements

Causes of variation (frequency)	
Human errors(17)	Uncontrolled variables(11)
• Strength of releasing the ball(8) • Releasing the ball at an angle(2) • Errors in measurement e.g. height, parallax, reaction time(7)	• Environmental factors e.g. wind(6) • Spot where the ball landed e.g. uneven(3) • Properties of the ball e.g. loss of elasticity, wearing out(2)

As in P1I1 and P1Q1 (see Sections 4.4.1 and 5.4.3 respectively), human errors were mostly cited as the cause of variation. The combined responses for human errors was 60.7% (17), almost similar to the same probe in P1I1. In fact, almost half the PSTs cited *only* human errors in their responses, ignoring other sources of errors; the response below was typical of the rest:

> I think it should vary because of human error. The difference should be slight if the process is done carefully. (I2-35)

On whether the PSTs would take more than ten readings, 60% (15) responded they would not, and several among them thought they should stop at a fixed number of readings (five or ten being the most popular choices):

> I would stop at five…five times is the best because after that it's a waste of time. (I2-50)

> For an experiment, ten times of repeated measurements is just enough. I think ten times is just right. (I2-46)

Their decisions were not influenced by the quality of their repeated data rather by their practical concerns influenced by their prior educational experiences:

> In the classroom context, most of the time, we are asked to do three times. If we are doing this for a real purpose, I would do ten or twenty times. (I2-35)

A number of PSTs indicated they would not go beyond ten readings if they had already obtained consistent data, but if they did not, they would likely continue until their goal was met; the response below underscored this notion:

> I'll take more data after bounce number 10 because there's still variation, the highest being 41 and the lowest is 38. (I2-9)

As explained in Section 2.2.6, the standard deviation or the standard error would be the appropriate indicators to guide the number of repeats to be taken but neither of these concepts were ever mentioned. Nevertheless, five PSTs expressed they would base their decisions on the degree of variation. They suggested ten readings would be enough if the variation was small and deemed sufficient to give a reliable mean value. They did not elaborate how they determined the degree of variation, but the evidence seemed to point to common sense.

The final question concerned when to stop taking repeats. The derived categories were as follows:

(a) Close readings(7);

(b) Consistent readings(12);

(c) Enough for a mean calculation(3);

(d) A fixed number of repeats (3).

The frequencies of (a) and (b) together accounted for 76%, and essentially, the PSTs would stop taking measurements once they reached zero or a small degree of variation (in their view). A response from (a) is given below:

> It depends on the variance, I'll do 5 to 8 times and check the variance and I'll stop if it's small but I'll do another 10 if it's big. (I2-52)

For PSTs in category (b) who claimed to be guided by consistent results, quite likely that all except three were not be looking for a true value since they had initially filled their table with varied data (from Question 7). For this category of PSTs, their search for a consistent reading could either be an overstated goal or

they were just highly confident of being able to achieve similar rebound height measurements (due to high repeatability of distance measurements from their past learning experiences). The three PSTs in (c) could have also been lumped together with those that looked at variation (the 76%) if not because of their stated goal of finding a mean value. All three had a similar response like the one given below:

> I would look at these 10 readings, if the *difference is quite small* and the average is around there, I wouldn't waste my time taking more readings. (I2-43)

Finally, a small group of three PSTs expressed they would stop at a fixed number of repeats. Their decisions were very much guided by the routines developed in their prior learning experiences:

> We've been trained from the few years of secondary education. I think it's always just between 5 to 6 readings (I2-53).

Question 8:

If they were to report the rebound height of the rubber ball as a measure of its 'bounciness', Ask how they would go about doing it, how they arrived at the value, and why that value?

Eighteen PSTs chose to report a mean value. Within this group of PSTs, a large sub-group indicated they would use *all* their repeats to calculate the mean value:

> I would probably take all ten values as not taking some might mean I was bias. (I2-31)

If there were anomalous results, they would replace them first:

> I should respect all the values I had taken so I would take all the values. If one or two were very different, I would [just] do those again. (I2-43)

A smaller sub-group consisting of only four PSTs would give the mean and the range; one PST among them even suggested the SD:

> I would find the standard deviation. I shall report like 40 ± SD; I'll report a value that is between 39.5 and 40.5. (I2-37)

This exceptional response could be based on a routine idea since none of the PSTs ever mentioned SD in their earlier responses.

The remaining seven PSTs chose to report the bounciness using the mode (a recurring value that appeared most number of times):

> I'll take a value that has been repeated many times as it should be accurate. (I2-38)

Question 9:

Two students **A** and **B** carried out the same investigation with different rubber balls of the same brand and reported the results shown in the table below.

Bounce number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Student **A** Bounce height (cm)	40	45	36	50	39	42	42	42
Student **B** Bounce height (cm)	45	44	43	42	39	39		

Both students **A** and **B** reported bounce height of 42 cm by adding all the rebound heights and dividing by the total number of bounce for each. With whom, **A** or **B**, do you most closely agree? Explain your choice.

Based on SD calculation, **A** = 3.905cm and **B** = 2.309cm, which meant the spread in **A**'s readings was greater than **B**'s, but this information was not provided to the PSTs.

Twelve PSTs chose **A** over **B,** and all except two based their decisions on **A** having more repeats than **B**. The PSTs generally believed that more data points were "better" since the mean value would be more "accurate"[29] given it was derived from more repeats:

> He (Student **A**] has more data and **B** has only six. I feel it's accurate because of more data and the average of; the more the better (I2-40).

Several PSTs who chose **A** provided a second reason in that a consistent value of 42cm appeared in the last three measurements:

> **A** because he did more times, and finally [in] the last few, he got the same results. (I2-34)

This idea of consistent repeats was in fact the only one that guided two PSTs to choose **A**. Thirteen PSTs chose **B** over **A,** because of the smaller variation in **B**'s readings:

> I agree with **B** because although the number of tries is lesser, the difference between the values is not [as] large as Student **A**. (I2-30)

[29] The PSTs meant reliability here.

A number of PSTs justified their choice by citing **A**'s larger range of readings:

> The biggest variance here [student **B**] is 4 whereas for Student **A**, he has a result of 50 which is 8cm more than the mean; and at bounce number 3, he has a 36, which is 6cm away from 42. (I2-52)

Another reason that was used by seven PSTs was based on the abnormal reading (in their view) in **A**'s measurements:

> Because for student **A**, there is a large discrepancy at the fourth bounce which is 50 and it differs greatly from the other bounces. (I2-29)

Section 5.4.1 in P1Q1 showed that many PSTs thought it was only right to compare data sets with equal number of repeats. The same idea of "fairness" was revealed in the responses given by three individuals; one is shown below:

> Since Student **B** has done it only six times, I would consider Student **A** have done six times [as well], ignoring [bounce number] 7 and 8. (I2-29)

Question 10:

> Instead of measuring the first bound height by eye against a metre rule, the students used a *self-automated* digital camera to record the first bound height.
>
> Would they obtain similar results (as varied as before)? In your opinion, how many repeats should be appropriately conducted (as many, or more, or less than using the eye)?

Three PSTs claimed the readings would remain as varied as before because human errors could not be discounted on the basis of the significant involvement of the investigator in the task:

> It will be varied like there will still be a difference, 40, 41cm as they [*investigators*] are still the ones who dropped the ball. (I2-34)

Most of the other PSTs responded the variation would change and become smaller. A typical response is shown below.

> With the digital camera, [the readings] would be more similar; the readings will be less varied…you have taken out the human factor and used a digital camera, which was more precise. (I2-37)

Many PSTs like the one below did relate to the "*self-automation"* as the reason for the reduction in the degree of variation:

> The value I get is more accurate because the digital camera is self-automated and does not have human error. (I2-42)

19 PSTs claimed *more* of the same reading say 40cm could be obtained because of "the higher accuracy of the digital camera as compared to the human organ, which is the eye" (I2-29). But it would be impossible to obtain a similar reading *throughout* because the task was still opened to human errors:

> Although we have the digital camera to capture the height, I think the strength in which the student is releasing the ball is not constant. (I2-47)

Two PSTs claimed it was possible to achieve a "perfect" measurement, but this was significantly fewer than the number of PSTs who believed a "perfect" instrument or measurement existed in P1I1 (see Section 4.4.2). Earlier in Question 7, three PSTs claimed the data would remain the same throughout with the use of the eye only; for the current probe, they still maintained their stance, and two added the measurements would now have more decimal places. The response below came from one of them:

> The same results, but more precise, maybe to two decimal places because now they use a machine. (I2-50)

The last question explored the PSTs' ideas about the number of repeats in the light of using a more reliable device (a digital camera). The responses were rather diverse; nevertheless, after coding analysis, the PSTs could be differentiated into three categories. Eleven PSTs claimed they should perform the number of repeats equal to the number when using the eye only. Most of these PSTs seemed to think it would only be "fair" to compare data sets only if they have the same number of repeats, as shown in the following response:

> I just do the same number as with the eye when we needed to do a comparison. (I2-32)

Another five PSTs chose to take less repeats; one response from this category is shown below:

> [*Since*] the digital camera is more accurate than the eye, ten will be more than enough but five would be just sufficient. (I2-51)

Basically, their reasons stemmed from the idea of instrumental reliability:

> The digital camera usually doesn't make much mistake [errors] because it is more trustworthy and reliable. (I2-33)

Finally, a third category of PSTs claimed they would attempt a fixed number of repeats. Many suggested ten repeats would be just right but a few suggested less, for instance, five repeats:

> I would go with five readings…it's the norm. (I2-53)

Review of Probe 7

In Question 7, the imaginary data entered in the table helped the PSTs to craft their responses when they needed to explain their ideas concerning the number of repeats and when to stop repeating measurements. The table also allowed me to check the notion of true values amongst the PSTs who brought up the idea of consistent repeats.

A number of findings from the probe concerning errors can be highlighted. The fact that the PSTs had a tendency to pin down human errors as the main cause of variation was again established, and we could see its strength as a source of error despite introducing the self-automated digital camera in Question 10. Another notable observation was the sources of errors due to uncontrolled factors given in Question 7; the descriptions were "richer" and the sources more diverse. This might be due to the modifications made to the probe in terms of its description and diagrammatic depiction, which could have helped the PSTs to visualise the investigation better.

The evidence to Question 8 showed a common characteristic among the PSTs in adhering to certain routine ideas. For instance, the three PSTs who claimed the same data would be obtained for all ten readings using the eye or the digital camera still insisted in calculating a mean value to arrive at the "bounciness" value because it was a "ritual" they needed to follow. Another

routine idea that appeared frequently was the fixed number of repeats to be taken. The same idea led some PSTs to insist that they should only compare data sets with equal number of repeats.

In Question 9, we could see several ideas linked to the characteristics of a data set "competing" with each other to indicate which data set would be more reliable. However, ideas like standard deviation and standard errors seemed to be completely absent, thus affirming earlier observation from Section 4.4.2 that the PSTs were likely to be unaware of the statistical concepts.

6.4.4 Probe 8: The "Pendulum" probe

Probe 8 was based on a well-known physics investigation to find the period of a pendulum. From my knowledge of the local science curricula, the PSTs would have carried out the investigation at least once in their previous educational settings.

A student wanted to see how the length of a pendulum affects the period T, which is defined as the time taken for one complete swing. She took the time taken using a stopwatch for the pendulum to swing to-and-fro twenty times. She repeated the same measurement at different lengths three times. The results were recorded as shown in the table below.

Length of pendulum in cm	Time take for 20 swings in seconds		
	1	2	3
40	26	25	28
60	31	32	32
80	36	36	41
100	40	47	41
120	50	43	44

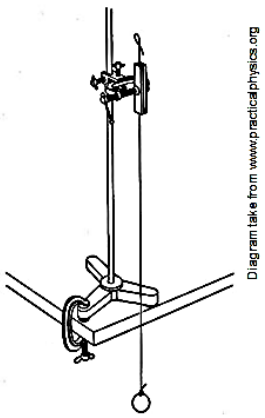


Diagram take from www.practicalphysics.org

The accompanying questions will be presented separately

The presented table in the probe consisted of a continuous IV (length of pendulum) against a continuous DV (time for twenty oscillations). The PSTs would need to process and interpret the data to draw out the proportional relationship between the length of the pendulum and the time taken for twenty

oscillations, but these might be hindered by a number of “counterintuitive” data (encircled in the figure above, but not indicated to PSTs).

Specific Objectives

Based on Section 4.4.4, the PSTs would either read the data across the table (the “rowers”) or down the columns (the “columners”). The actions taken by the PSTs to process the tabulated data would point out which category they belonged to (i.e.; if they inspected the three repeats, and then calculated a mean value, they would probably be “rowers”, but if they examined the timings down the columns against the lengths of the pendulum, then they would likely be the “columners”). Based on Section 4.4.4, the “rowers” might either ignore the “counterintuitive” data in the repeats or replace them before calculating a mean value. The “columners” might suggest either carrying out a whole trial again or replace specific data that disrupted the trend.

Results and Discussions

Question 11:

Describe in detail what the results in the table show you. Are there concerns with any of the data shown in the table? If you were the student, how would you proceed with the experiment?

From the outset, the PSTs seemed to recognise the relationship between the length of the pendulum and the time for twenty oscillations. This implied a strong familiarity with the task, but it also posed a challenge to categorising the PSTs as “rowers” or “columners”. When the responses were coded and analysed, the results suggested a high number of “columners”, but this could have resulted from the PSTs adopting a “columner’s” frame to examine the data in order to find the relationship between the variables.

Nevertheless, the analysed evidence did show evidence of both categories. The two categories are shown below, differentiated by the PSTs' main reasons and supported by an example each.

(a) **"Rowers",** *inconsistent* values across the rows(4):

> The results vary and they are not consistent in the sense that they are not the same; they vary by one or two seconds. (I2-35)

(b) **"Rowers",** *large range in the repeats taken at the same IV interval*(3):

> For 100cm, the range is a bit too much; the difference is about seven seconds when the rest was only about five seconds. (I2-51)

(c) **"Columners",** *absence of increasing trend*(8):

> 41 seconds occurred twice at 80 and 100cm; I am not concerned with the other readings. (I2-53)

(d) **"Columners"**, *irregular* trend between the IV and DV(8):

> I observed when the length of the pendulum was at 40, the change was not that much, but when it got to 100 or 120, the changes were quite a lot, 40 and 47 as well as 50 and 43. (I2-46)

On the question how they would proceed with the given data, most "rowers" suggested finding the mean value for all three repeats or taking more repeats to replace the "counterintuitive" readings only. To illustrate, an example of a "rower's" response is shown below:

> I would rather do all the timings for each length to be consistent. I have to do four times for each length so that the averages will all be based on four readings. (I2-48)

The response also showed the decision taken by 12-48 was influenced by their own flawed idea of "fair" (which in this case was the need to take equal number of readings for all IV intervals).

The sixteen "columners" suggested either to take a whole series of DV readings at every IV interval or to repeat certain DV values that seemed to be disrupting the trend (for e.g., those corresponding to 100 and 120cm lengths).

Review of Probe 8

The use of a familiar task posed a challenge to discovering how the PSTs handled data in the probe because the PSTs referred to their prior knowledge about the task to guide them in processing and interpreting the data.

Two PSTs displayed a pattern of thinking that showed ideas from both "rowers" and "columners", for instance I2-48:

> The longer the pendulum, the longer the time it takes for the pendulum to complete 20 swings. The results vary at each length; the time is not exact: the time taken for 40cm in the first, second and third [repeat] varies. (I2-48)

They might belong to a third but small category of PSTs who processed data by being both a "rower" and a "columner".

6.4.5 Probe 9: The "Osmosis" probe

This was the same probe from P1I1 (see Section 4.4.4) that first revealed the "rowers" and "columners". In P1I2, the "osmosis" probe was optional and used only if there was time during the interviews. At the end of the study, a total of eleven PSTs (about 44%) were interviewed using the probe.

A scientist wanted to see how temperature affected osmosis in a potato. She puts equal size chips in a 1 mol/dm^3 sugar solution at different temperatures and measured the change in mass of the 'chips' on a top pan balance and recorded the results as a percentage of the original mass after 4 hours. She repeated all the readings 3 times.

Percentage of original mass			
Temperature (°C)	1	2	3
15	83.46	78.88	80.91
30	77.5	82.67	80.59
45	82.66	65.47	74.24
60	67.76	93.32	73.18

Question 12: What does the data show? Explain all your reasoning?

Specific objectives

Since this probe had not been modified, its specific objectives would be similar to those given in Section 4.4.4. The data gathered from this study[30]

[30] A preliminary scan of the interviewees' transcripts indicated the existence of the two categories.

would serve to substantiate the P1I1 findings, and provide more information regarding the distribution of “rowers” and “columners” in the group of PSTs.

Results and Findings

From the analysis of the coded responses, seven PSTs could be classified as “columners”, four as “rowers”, and none from the mixed group. The differences in numbers between the categories, however, were too small to be of any significance. Nevertheless, a response each from the “columners” and “rowers” is given below to illustrate how the data had been processed by members of each category.

“Columners” (7):

> As the temperature increases…the length didn’t change much from 15 to 60^{o}C. I’m looking at the lengths of the chips comparing [them] at 15^{o}C and 60^{o}C. (I2-31)

“Rowers” (4):

> I am looking across... When the temperature increases, there isn’t any effect on the data across…I wanted to check whether the length of the chip in the first is consistent with the second and the third trial. (I2-50)

Cross-checking against the evidence given in the previous “pendulum” probe, the results showed that with the exception of one PST, the other three PSTs identified as “rowers” in this probe were classified as one in the previous probe. Likewise for the “columners”, with the exception of one PST, the other six PSTs were also identified as “columners” in the previous probe. The results implied the way PSTs processed (unfamiliar) tabulated data could be relatively stable across different tasks. But more studies must be done to verify this.

Review of Probe 9

In addition to the review done earlier in Section 4.4.4, a few other observations were noted. First, a number of PSTs were initially stumped by the question (after given some time to process the data) and did not respond immediately implying they did not have any prior knowledge about the

investigation. Second, the two decimal places in the given data might not be necessary and could have even stressed the PSTs; the decimal places could be removed if the probe were to be used again.

6.5 Conclusion of Phase 1

Chapter 6 generally re-affirmed many of the observations seen earlier in previous P1 studies that addressed Research Aim 1. One important but noteworthy observation was the frequent use of routine knowledge to decide on procedural actions.

After three studies which included several iterations of exploring the same procedural concepts, sufficient evidence and knowledge had been gathered to answer Research Aim 1 and develop a questionnaire as stipulated in Research Aim 2.

CHAPTER 7

PHASE 2 QUESTIONNAIRE 2 STUDY

7.1 Introduction to Phase 2 and Chapter Overview

After three rounds of studies in P1 geared towards establishing a "neutral ground" and developing "neutral tasks", the research reached a stage where it could move to the next phase of its study and focus on Research Aim 2. Chapter 7 will report on the next Phase 2 (P2) of this study; it will describe the first steps that are taken in the development of the questionnaire, one that can provide a quick and efficient way of getting evidence of PSTs' understanding of uncertainty in measurements and serve to inform the planning of teacher preparatory programmes.

In Section 2.5, the thesis proposed that understanding uncertainty in different measurements as specified in Section 1.7 would involve applying procedural understanding, which is defined in this thesis as the Concepts of Evidence (CofEv). This was largely investigated in P1; the gathered evidence would now be used to develop probes for the questionnaire and then to analyse them in order to achieve Research Aim 2. The P1 contributions to each probe in the proposed questionnaire will be distilled and stated together with the aim of the probe in the next section. This will then be followed by the results and discussions with the objective of improving the probes. The final section will present a summary of the chapter. Following the practice in past chapters, the current study shall be referred to as "P2Q2".

7.2 Structure of P2Q2 Questionnaire

With reference to the P2Q2 Questionnaire in Annex 3.4, Table 7.1 below specifies its structure with respect to the number of probes for different kinds of measurement and the aim of each probe.

Table 7.1 Structure of P2Q2 Questionnaire

Measurement of…	Probe	Title	Aims
a single datum	1(a) to (i)	The Instruments Test	The choice within a range of instruments that measure the same quantity but with different characteristics such as the resolution of scale, limits of detection, analogue/digital display, etc.
repeated data (of a DV related to a categoric IV)	3(a) to (c)	The Sole Test	To determine the surface that needed the most and least force to pull a shoe along.
	4(a) to (c)	The Bouncing Rubber Ball Test	To find the bounciest ball from two data sets.
repeated data (of a DV related to a continuous IV	2(a) to (f)	Repeats	To determine general procedural ideas about repeats.
	5(a)	Starting an Investigation	The sequence to take for the first four measurements in an investigation.
	5(b)	What next in an investigation	The next two measurements to be taken after "messy"[31] data were obtained in an investigation.

7.3 How were the questions asked?

The reviews in the P1 studies provided several critical insights towards the design of the questionnaire. As a result, several features were adopted in the P2Q2 instrument and these were:

(a) The language was kept simple. The stem of each probe was written using either short sentences containing one to two clauses or questions asking for a single idea.

(b) The frequent use of diagrams to illustrate and tables to organise data. This helped the PSTs to rely less on memory; thus, preventing information overload. The easier questions were generally placed first to motivate the PSTs towards completing the probe.

(c) Procedural concepts like human errors and mean values that seemed to be well understood by most, if not all PSTs, were not assessed. The probes looked beyond these concepts to explore the "thinking behind the doing".

[31]As in P1, "messy" here meant "high uncertainties" shown by a large degree of variation in the repeats and the absence of trend relationship between the variables.

(d) The questions were designed to provide the PSTs with an easy and fast method of indicating their answers without having to think much about how to articulate them. The latter was achieved by deploying multiple-choice questions that allowed the PSTs to select an answer from choices based on categories derived from P1 results that best fitted their understandings. Since the responses were easily classified, the analysis became straightforward. However, this did not mean the PSTs were unable to express their ideas freely as each multiple-choice question was accompanied by an open-ended question that allowed the PSTs to explain their answer. The response was expected to be brief and focused on one or two ideas at the most, thus allowing it to be easily recorded, coded, and analysed quantitatively.

7.4 What questions were asked?

This section links the contribution of the relevant P1 studies to the construction of the P2Q2 probes and describes the CofEv that underpinned each probe. The discussions will follow the order shown in Table 7.1.

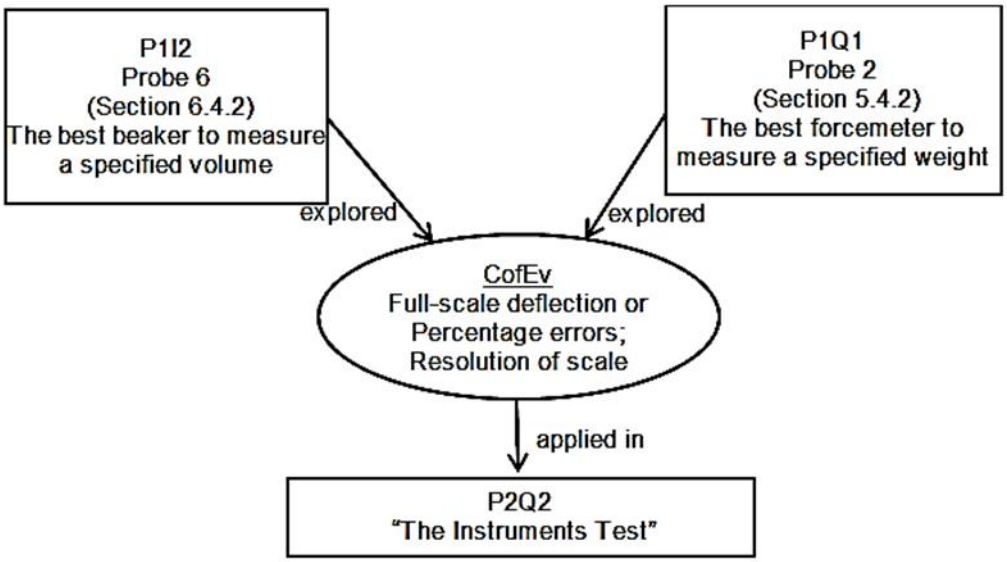
7.4.1 Probes for "a single datum"

The uncertainty in a single measurement may be concerned with the reading error associated with the limitation of a measuring instrument in terms of its resolution of scale. Thus, if a given quantity falls within the scales of two measuring instruments, the one that has a smaller resolution (and therefore, a smaller reading error) will give a more accurate measurement.

Choosing an instrument that "fits the purpose" may also lead to lower uncertainty (a sensible choice would be an instrument that could give the lowest percentage reading error, i.e., the quantity to be measured would be nearer the end of its scale).

The P1 studies informed the construction of the “The Instruments Test” conceptually in two areas. First, the choice of an instrument would depend on the instrument that could give the most *accurate* datum. Second was the CofEv that would be factored in making the choice of instruments (see Figure 7.1).

Figure 7.1 The development of “The Instruments Test” based on P1 studies



The “Instruments Test” sought to explore the PSTs’ understandings of uncertainty by looking at several measuring instruments commonly used in primary science investigations such as rulers, forcemeters, thermometers, voltmeters, beakers and measuring cylinders. All these instruments were equipped with a static-display scale. Recently, instruments with digital readouts have also been introduced to primary schools, so digital weighing balances and clocks were also included in the test.

In probes based on instruments with a static scale such as measuring cylinders and voltmeters, the PSTs were expected to decide their choice of instrument based on the concepts of full-scale deflection (FSD), percentage errors, and the resolution of scale. However, in instruments like the ruler, measurements could be obtained directly by reading off their static scale; thus, quite likely the PSTs would have to deal with only the resolution of scale. As for

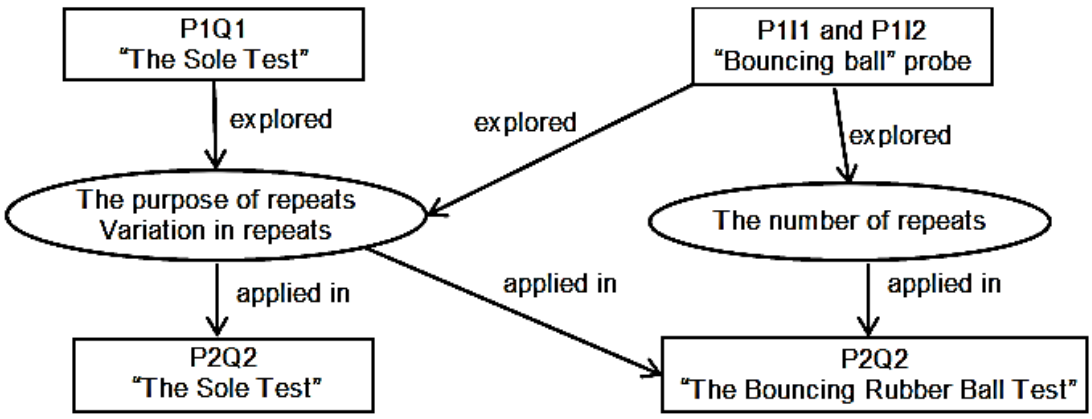
digital weighing balances, the PSTs might need to deal with their limits of detection (i.e., the maximum and minimum quantity that could reliably be measured), and the "scale readability" (i.e., the smallest change in mass that corresponded to a change in the displayed value, and usually expressed, for instance, as "up to 0.01g", etc.)[32].

The final probe dealt with the use of an instrument like the clock that has either analogue or digital display. The PSTs might misconceive a digital-display clock to be more accurate because of it modern and sophisticated appearance. Such a notion was drawn from a few responses to the "Bouncing Ball" probe in P1 (see Section 6.4.3) that claimed a digital camera could "capture" perfectly the height measurements.

7.4.2 Probes for "a data set"

The P2Q2 probes, "The Sole Test" and "The Bouncing Rubber Ball Test", used for exploring the PSTs' understanding of uncertainty in a set of repeats, were both modifications of P1 probes (see Sections 4.4.2, 5.4.3 and 6.4.3). The results obtained from the investigations of several key ideas explored in P1 contributed to the construction of these probes (see Figure 7.2).

Figure 7.2 The development of "The Sole Test" from P1 studies

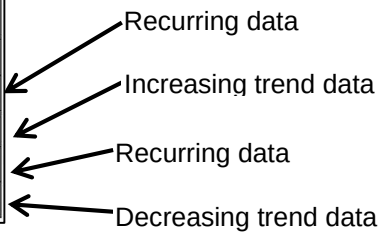


[32]To illustrate, 2.358g weighed on a scale with 0.001g readability would read "2.358g", but on a scale with 0.01g readability, the display should show "2.36g" (American Weigh Scales, 2011)

"The Sole Test" (that investigated how different soles could affect the "slippiness" of a shoe) in P2Q2 had been modified to replace the human investigator with a robot pulling the shoe with a forcemeter (see Annex 3.4). This was to avoid the PSTs from making references to human errors as the cause of variation. Additionally, the data had been modified to display certain characteristics that could influence the PSTs' in their selection of data for the most and the least pulling force (see Figure 7.3). Such data characteristics that represent patterns of reliable data to the PSTs were based on evidence derived from P1 studies (see Section 5.4.3).

Figure 7.3 Data characteristics in "The Sole Test"

Type of Surface	*Pull force (Newtons)*		
	1	*2*	*3*
Soil on the school's playground	10	10	16
Grass on the school field	10	12	14
Carpet in the school's library	4	6	6
Wooden floor in the school hall	7	6	3



The next probe, "The Bouncing Rubber Ball Test", studied the PSTs' decisions on the number of repeats to be taken when comparing two sets of repeated DV measurements of re-bounce heights taken for different IV categories (for e.g., Ball **A** and **B**). The number of repeats will be decided based on the difference between the mean heights of each data set in relation to the variation in the repeated DV measurements of both sets.

Four situations were deemed possible (see Table 7.2) but only the first three were studied. The fourth situation (d) was not posed because it required a large number of repeats that might not fit into the questionnaire; besides, it could be too taxing (and time consuming) for the PSTs to look at a very large number of data. Additionally, the objective of the probe to check for understanding could be met adequately using the first three situations; the fourth situation was not critically needed.

Table 7.2 Four possible situations to choose the bounciest rubber ball[33]

Situation	Difference in mean heights of IV categories (e.g. Balls **A** and **B**) (1)	Variation in rebound heights (DV) of balls (2)	Difference between (1) and (2)	Amount of readings needed to be taken (suggested number)
(a)	Small	Small	Small (difficult to distinguish between balls)	Many (about 20)
(b)	Large	Small	Large (balls easily distinguished)	Few (about 3)
(c)	Large	Large	Small [given data allowed balls to be distinguished, but not as easy as (b)]	A reasonable number (about 10)
(d)	Small	Large	Large (very difficult to distinguish balls)	Very large number (above 20)

7.4.3 Probes for "DV data of a continuous IV"

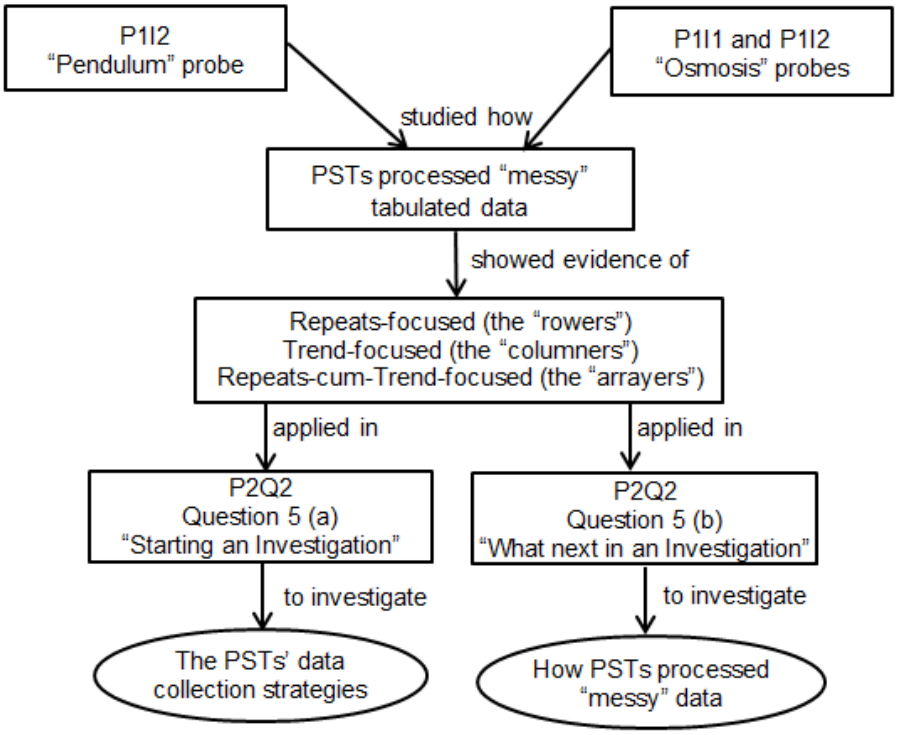
The first of two probes, Probe 4, that were used to investigate the PSTs' understanding of uncertainty in DV data related to a continuous IV was titled "Repeats". Its stem reads: "In an investigation, we often take repeated readings for each value of the independent variable" - this should prompt the PSTs to respond to the six statements with the context of a variable-based investigation in mind. The general statements were aimed at exploring the PSTs' tacit understanding of uncertainty that would guide their intuitive actions or decisions at different stages of an investigation from planning to performing measurements to processing and interpreting data.

The statements constructed for "Repeats" were refinements of the 14 statements from P1Q1 (see Section 5.4.1). After the P1Q1 review, several statements were discarded because they were either redundant or lacked clarity. The retained statements were then modified and checked for bias and ambiguity. The finalised statements looked at several CofEv: anomaly, random human errors, variation and number of repeats.

[33] The data used in Probe 4 were based on authentic data derived from a database at the University of Durham. Annex 7.1 gives more details for the 3 situations (a) to (c) in Table 7.2.

The final Probe 5 had not been tested in P1. Nonetheless, the ideas behind Probe 5 could be traced back to several P1 probes in Sections 4.4.4, 6.4.4, and 6.4.5 (see Figure 7.4).[34]

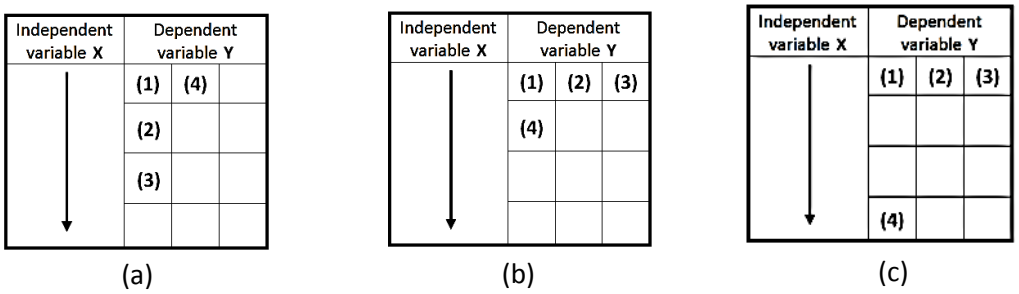
Figure 7.4 The development of Probe 5 from the P1 studies



Probe 5(a), titled "Starting an Investigation", was aimed at studying how PSTs would plan the sequence of their first four DV measurements in an investigation. Based on the CofEv, the best data collection plan should consider (a) the range of IV measurements that would reveal the full extent of the relationship between variables, and (b) the appropriate number of DV repeats to represent the degree of variation. Bearing these in mind, looking at Figure 7.5 where the number within parentheses represents the order in which the measurements could be taken, the best sequence for the first four measurements would be (c)[35]. The PST who picked (c) could be an "arrayer" ("Repeats-cum-Trend-focused").

[34] New terms introduced here are merely to help depict the "rowers", "columners", and "arrayers" more clearly.
[35] If the numbers were used as labels and not meant to show sequence, then the order could also be "4-1-2-3", "1-4-2-3", etc.

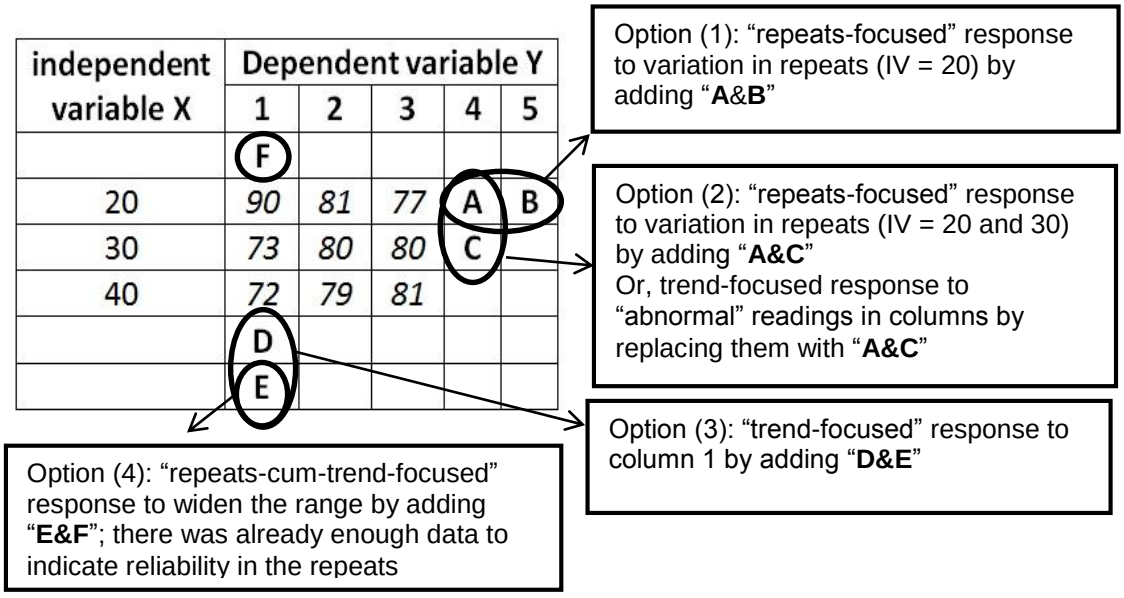
Figure 7.5 Probe 5(a)



The other two sequences (a) and (b) might appeal to a "columner" (trend-focused) and a "rower" (repeats-focused) respectively. Looking back at P1, the instruments that were used then were not able to reveal the "arrayers" conclusively; thus, Probe 5(a), and the next Probe 5(b) served as further probes to check if the category existed.

Probe 5(b), "What next in an Investigation", was designed based on P1 evidence that showed how "messy" data could be processed by a "rower" or a "columner". A "rower" being "repeats-focused" might take more repeats to address the uncertainties whereas a "columner" being "trend-focused" would take more data in order for a trend between the variables to emerge. Based on these notions, five sets of data were created; some were more "messy" in the rows than the columns or the other way round (see Figure 7.6.).

Figure 7.6 Data set of Probe 5b (i) and possible reasons for choosing Options 1 to 4



In Figure 7.6, the PSTs were asked to pick their next two readings from four given options: if a PST picked "**A&B**", the PST was likely a "rower"; if the PST picked "**D&E**", the PST could be a "columner"; and if the PST chose "**E&F**", he or she could be an "arrayer". The PSTs would have to respond to five different probes and it was deemed adequate to reliably identify a "rower", "columner" or an "arrayer". The last two sets of "messy" data were slightly different in that they allowed the PSTs to have a free choice of their next two readings from any of the data provided (the choices offered in the first three probes covered all possible categories; nevertheless, a freedom of choice would permit "nuanced" responses from any of the categories and to see how else the measurements might be carried out).

7.5 Results and Discussions

The results will be discussed according to the order shown in Table 7.1. By analysing and explaining the results, we would be able to study the efficacy of a probe especially with regards to it meeting its aim. In the discussions, suggestions to improve the probes will also be made based on the evidence obtained.

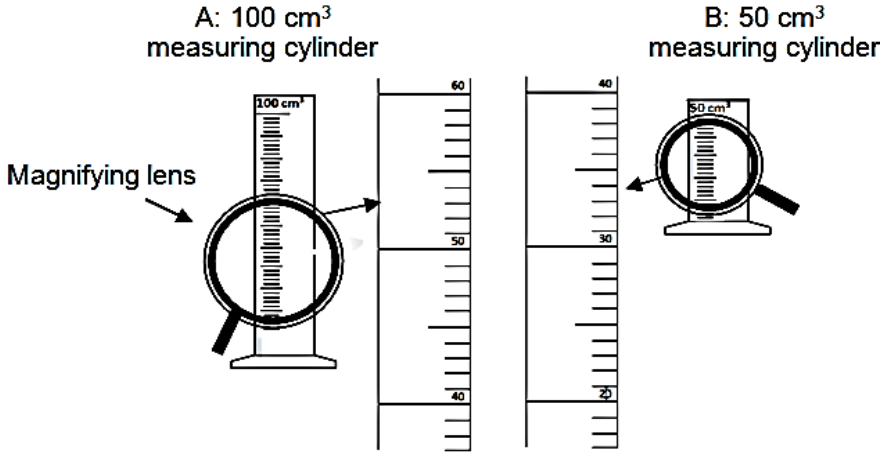
7.5.1 "The Instruments Test"

To focus on the understandings behind the choice of different types of measuring instruments (see Section 7.4.1), the results for "The Instruments Test" will be presented by "clustering" the probes according to the CofEv that underpinned them. In order to show how a probe in the test had been presented, Probe 1(e) will be used as an example (see Figure 7.7).

In Figure 7.7, two measuring cylinders of the same dimension but different capacities (i.e., 50cm^3 and 100cm^3) were given; both could actually be used to measure the given volume of 35cm^3 accurately as the resolution of

scale in both cylinders was the same (reading error = ±0.5cm^3). The best response, therefore, will be: "it doesn't matter which".

Figure 7.7 Probe 1(e): Choosing a measuring cylinder to measure 35cm^3 of solution



The results for the first cluster of items are reported in Table 7.3.

Table 7.3 Analysis of Probes 1(a), (b), (d), and (e) (N=20)[36]

1	Instrument	Choice	Number of responses (%)	FSD	Resolution of scale	Non-response[37]
(a)	Forcemeters	*idmw **(A)0-10N** (B) 0-25N (C)0-50N	3(15) **13(65)** 4(20) 0(0)	**4**	**8**	**1**
(b)	Beakers	idmw **(A)100cm^3** (B)150cm^3 (C)200cm^3	3(15) **13(65)** 4(20) 0(0)	**5**	**6**	**2**
(d)	Voltmeters	idmw **(A)0-6V** (B) 0-12V (C)0-18V	1(5) **7(35)** 11(55) 1(5)	**1**	**5**	**1**
(e)	Measuring cylinders	**idmw** (A)100cm^3 (B)50cm^3	**5(25)** 1(5) 14(70)	**0**	**5**	**0**

*idmw= "it doesn't matter which"

Drawing from Table 7.3, we could see for Probes 1(a) (forcemeters) and 1(b) (beakers), most PSTs agreed with the experts' responses, but not for 1(d) (voltmeters) and 1(e) (measuring cylinders). Table 7.3 also states the PSTs' CofEv for choosing the option, which informed us of the concepts that supported their decisions. It must be noted that all PSTs either gave FSD or the

[36] The experts' answers are emboldened. The same representation will be adopted for all other results tables.
[37] Includes vague responses

resolution of scale in their explanations, and none had given both. Additionally, none had also picked the experts' choice with a reason other than the two given (one was unacceptable because the explanation given was vague and incomprehensible).

Table 7.4 shows examples of the PSTs' explanations based on the two key concepts.

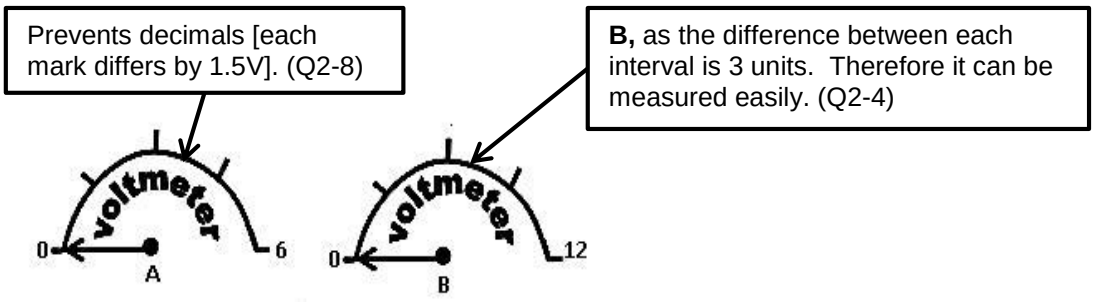
Table 7.4 Quotes from PSTs using FSD and the resolution of scale

1	FSD
(a)	10N [forcemeter] is closest to 8N. (Q2-15)
(b)	$80cm^3$ is closest to $100cm^3$ making the measurement less likely to have error. (Q2-14)
	Resolution of scale
(d)	The subdivision between each marking is the smallest allowing accurate measurement. (Q2-6)
(e)	Both cylinders have the same division, each division increases the reading by $1cm^3$, so it does not matter which as both would have the same accuracy. (Q2-3)

We could see from table 7.3 the concept of FSD was not as frequently applied as the resolution of scale, which reinforced earlier P1 observations in Sections 5.4.2 and 6.4.2. Those who picked the distractor "idmw" in Probes 1(a) and (b) were generally not concerned about getting an accurate reading whereas those who picked the distractor **B** generally believed the bigger scales could easily "accommodate" the specified quantities, which should fall somewhere in the middle of the scales.

To explain the results for Probes 1(d) and 1(e), the responses given to the strongest "distractor" were analysed. In 1(d), a number of PSTs chose **B** (over **A)** in order to avoid the odd-scale in voltmeter **A** (see Figure 7.8, together with examples of explanations why each option was picked). In order for this probe to work effectively, either the scale in **A** (divisions = 1.5V) or the specified quantity in the probe (i.e., 5V) has to be rectified.

Figure 7.8 Probe 1(d): why PSTs prefer B to A



As for Probe 1(e), despite the scales in both cylinders were shown to be the same dimension (see Figure 7.7), most PSTs(11) imagined the resolution of scale (and therefore, the reading error) in the 50cm^3 measuring cylinder was smaller than the 100cm^3, and therefore, was be more accurate. To improve, the diagram has to be enlarged and the volumes clearly marked. A small number (3) seemed to have opted for the 50cm^3 measuring cylinder based on FSD (35cm^3 was nearer the end of its scale), which implied the option **B** could potentially be a good distractor.

The second cluster of probes was based on measuring instruments that depended only on the resolution of scale. Table 7.5 gives the analysis of these probes.

Table 7.5 Analysis of Probes 1(c) and (f) (N=20)

1	Instrument	Choice	Number of responses (%)	Resolution of scale	Non-response
(c)	Thermometers	idmw (A) 0-50^{o}C (5^{o}C intervals) **(B) 0-50^{o}C** **(2.5^{o}C intervals**)	3(15) 0(0) **17(85)**	**16**	**1**
(f)	Rulers	**idmw** (A)10cm (0.1cm intervals) (B)1m (0.1cm intervals)	**5(25)** 15(75) 0(0)	**5**	**0**

As shown in Table 7.5, the results from Probe 1(c) showed the PSTs chose the thermometer with a smaller resolution of scale implying they understood the concept. However, the results from Probe 1(f) seemed to contradict this finding as most chose distractor **A** instead of “idmw” (which

implied the PSTs did not notice the two rulers had the same resolution of scale). Those who chose option **A** relied on practical reasons:

> The string was not long [8cm] so there was no need to use the metre rule. (Q2-17)
>
> It would be troublesome to use a long metre ruler which occupied a larger space on the workbench. (Q2-2)

Perhaps for future use, the instructions must be clearly stated in the question stem of Probe 1(f) in order for the PSTs to base their choice solely on procedural ideas instead of logistical considerations. I also contemplated whether the Probe 1(f) could be modified by using a 20cm (instead of 1m) ruler to measure the given length of 8cm but it might become an issue of test validity since such ruler was not a standard item in the local school laboratory.

The third cluster consisted of two probes based on digital weighing balances whose choices would be underpinned by two CofEv. The first of these would be the "limits of detection". The weights of the given objects [a pencil in Probe 1(g), and a fish in Probe 1(h)] should fall within the limits of the weighing balances presented in the table, however, some PSTs argued against these. Table 7.6 shows the analysis of results for the two probes.

Table 7.6 Analysis of Probes 1(g) and (h) (N=20)

1	Instrument	Choice	Number of responses (%)	Limits of detection	Scale readability	Non-response
(g)	Digital weighing balances	**idmw**	**12(60)**	**12**		**0**
		(A) 0-100g (reads to 0.1g)	7(35)			
		(B) 0-1000g (reads to 0.1g)	1(5)			
(h)		idmw	1(5)			
		(A) 0-1000g (reads to 0.01g)	5(25)			
		(B) 0-1000g (reads to 0.001g)	**14(70)**		**14**	**0**

The results show the majority of PSTs had no difficulty choosing the best balance based on the "limits of detection". Nevertheless, in Probe 1(g), seven PSTs believed a pencil should be rightfully measured by a 0-100g weighing

balance since it would be light enough (again, basing their choice on practical reasons). The following response typifies this group:

> A pencil won't weigh so much till 1000g. A 100g weighing balance would do. (Q2-14)

The second underlying concept was "scale readability" and was tested in Probe 1(h). 70% of the PSTs chose option **B**, which could read up to 0.001g (the experts' choice). However, five chose the other balance that read to 0.01g and questioned the need to weigh a fish accurately:

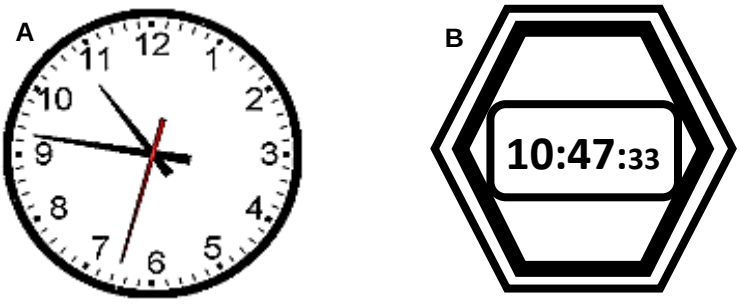
> [**B** if] greater accuracy was needed for scientific purpose. If normal daily purposes, use **A.** (Q2-11)

> For accuracy sake, choose **B** but if it's meant for the market, **A** will be better. (Q2-18)

The five PSTs who opted **A** probably had different contexts in mind. This affirmed earlier P1 observations about the influence of the "perceived purpose of measurement" (Lubben et al., 2004) on the PSTs' choice of instruments.

Figure 7.9 belongs to the final Probe 1(i), and shows the images of an analogue and a digital clock. In terms of "scale readability", both were equal right down to the seconds. The question was which clock would give the best (i.e., most accurate) time.

Figure 7.9 Probe 1(i): Analogue versus Digital clock



The experts' choice was "idmw" if it was based on an everyday context of telling the time, but a "digital clock" if it was to be used for a pendulum experiment since the analogue clock was known to give parallax errors or reading errors if the second hand moved continuously.

Table 7.7 shows the analysis of results.

Table 7.7 Analysis of Probe 1(i) (N=20)

1	Instrument	Choice	Number of responses (%)
(i)	Clocks	idmw	5(25)
		(A)analogue clock	0(0)
		(B)digital clock	15(75)

The analysis showed a clear preference of digital over analogue clocks. Two of the five PSTs who chose "idmw" based their decisions on the fact that both clocks could read up to one second:

> Both choices can tell time down to the seconds (Q2-3)
>
> Both clocks read hours, minutes and seconds (Q2-5)

Another two chose "idmw" because they could not tell the purpose of telling the time as required by the question; one of the responses is shown below:

> For science experiments, **B** would be good. On normal days, both clocks would do. (Q2-4)

Thus, to avoid a similar situation in the future, Probe 1(i) needs to be modified to have two questions; one based on everyday context, and another for use in an investigation.

For the fifteen other PSTs who chose digital clock, their explanations could be grouped into three categories (incidentally, each category had five PSTs). The first found it was easier to tell the time from a digital clock because of its clearer display:

> It [digital clock] states everything explicitly in numbers and it is clear. (Q2-10)

This again shows the PSTs falling back on practical reasons that should be avoided in the future if proper instructions were to be given.

The second claimed the analogue clock was prone to parallax errors:

> For clock **A** [analogue], there is a chance that there could be parallax error when reading time. (Q2-15)

The reason was acceptable if the PSTs had in mind the context of a pendulum experiment.

The last category, however, seemed to believe the digital clock was more accurate than the analogue as conveyed by the following quotes:

> Greater accuracy in seconds…. (Q2-11)
>
> It is more accurate down to the seconds. (Q2-13)

Their understanding of the accuracy of the digital clock compared to analogue clock could have been distorted by a misconception influenced by the outer appearance of the clocks. It would be interesting to see how this group of five PSTs would have chosen between a digital clock and an analogue clock with a slightly better accuracy.

Overall, the probes in "The Instruments Test" showed the PSTs had tacit understanding of choosing the best measuring instruments to measure a specific quantity, but seemed to have some difficulties articulating the underlying concepts. In Probes 1(a) and (b), the PSTs relied mainly on the concept of resolution of scale rather than the ideas of FSD or percentage errors. The reason could be due to the lack of knowledge of the concepts, which has implications for teaching and learning. As for digital instrument, the PSTs did not seem to have any difficulty choosing the best balance based on the "limits of detection" and they were generally aware that those capable of giving more significant numbers in their measurements with were more accurate. In improving the probes, instructions can be added to question stems of certain probes in order to prevent the PSTs from giving practical rather than conceptual reasons or by having clearer diagrams (e.g., the measuring cylinders can be enlarged and scales clearly marked) or to clearly state the measurement context (e.g. the fish is measured for a science investigation) or its purpose (e.g., the clocks are used to measure the period of a pendulum).

7.5.2 "The Sole Test"

The results obtained for Probe 3(a) are presented in Table 7.8 below:

Table 7.8 Why each surface was tested more than once? (N = 20)

Options		Number of responses (%)
A	to check the first reading	6(30)
B	***to see how much the readings vary***	10(50)
C	to get the same reading a few times	3(15)
D	to practice and get better	1(5)

50% seemed to have the right understanding for taking repeated readings but a reasonable number (30%) also thought a reading was repeated to check the first reading. The most intriguing outcome, however, was that only 15% chose option **C**, which was significantly less than what were found in P1 earlier. It could be the PSTs in P2Q2 had been taught explicitly that the purpose of repeats was not about getting the *same* results (which also implied the majority were not looking for true values). If access was granted, a follow-up interview with the PSTs would have established the reason for the low response. Nevertheless, the insignificant number of subjects in the sample meant the results could not be generalised.

Probe 3(b) was based on Figure 7.3 shown earlier on page 190. The PSTs were asked to state which surface required the *most* and *least* pull force. The pair of data sets to be compared was deliberately set up to give the same mean value; the PSTs therefore were unable to decide by giving a rote response using the value. As described in Section 7.4.2, each data set had certain characteristics (for e.g., increasing, etc.) that could influence the PSTs to select them. Because of the small degree of variation in the data sets and the insufficient readings to distinguish them, the experts chose option **C**, "I cannot tell which", for both questions. The results for the "most" pulling force revealed that option **C** was indeed the choice for 50% of the PSTs (see Table 7.9).

Table 7.9 Reasons for the "most" pulling force (N=20)

Options	Reasoning ideas	Number (%)	Quotes
A Grass	Substantive knowledge	1	The grass may cause it to be harder to pull the shoe along, more friction. (Q2-4)
	Anomaly in data	3	Because of the third reading on the surface of soil, it differs greatly from the first two readings. Hence it may be an ambiguity [anomaly]. (Q2-2)
	Total (%)	4(20)	
B Playground	Highest reading	4	The largest pull force [16] was used to pull the shoe along on the soil on the school's playground. (Q2-1)
	Substantive knowledge	2	There are more contacts, and it is a rougher surface compared to grass; so, there is more friction for the sole to overcome. (Q2-16)
	Total (%)	6(30)	
C I cannot tell which	Mean value	8	When averaging out the pull force between soil and grass, the pull force amounts to 12N for both. (Q2-7)
	Variation	2	Both options **A** and **B** have about the same differences. (Q2-11)
	Total (%)	10(50)	

The reasons for the PSTs' choices were analysed and categorised (see "Reasoning ideas" in Table 7.9). Overall, the data characteristics did not seem to strongly influence the PSTs' response; a few noted the third reading in "Playground" as the highest pulling force, but there were others who saw it as an anomaly. Several PSTs ignored the data completely and chose instead to base their decisions on their own substantive knowledge. A number chose option **C** as they could not tell the difference between the data sets based on the mean value, and not because there were variations in the repeats and/or insufficient readings.

Similar patterns of response were observed in the responses to the question about the surface that gave the "least" pull force. As shown in Table 7.10, only two gave acceptable reasons based on the variations in the data sets that were compared. In terms of data characteristics, the decreasing trend plus the presence of the lowest reading in the data set for "Wooden Floor" seemed to have influenced six PSTs to select it.

Table 7.10 Reasons for the "least" pulling force (N= 20)

Options	Reasoning ideas	Number (%)	Quotes
A Carpet	Anomaly in data	2	The third value taken at the wooden floor appears to be an anomaly as it varies largely from the first two values. (Q2-19)
	Total (%)	2(10)	
B Wooden floor	Substantive knowledge	1	Wooden floor because it does not have that much friction compared to the carpet. (Q2-4)
	Decreasing trend	6	The readings decreased throughout the attempts. (Q2-9)
	Total (%)	7(35)	
C I cannot tell which	Mean value	7	Their averages are the same value. (Q2-7)
	Substantive knowledge	2	A wooden floor in the hall should be smooth whereas the carpet should be the rough one. (Q2-2)
	Variation	2	The reading for the wooden floor varied too much to make a sound/logical comparison (Q2-16)
	Total (%)	11(55)	

Probe 3(c) further extended the idea of different data characteristic influencing the PSTs' follow-up actions or conclusions (see Table 7.11).

Table 7.11 Data Table for Probe 3(c)

Location of surface	Pull force (Newtons)						
	1	2	3	4	5	6	7
(I) Classroom	14	11	12	13	10		
(II) HOD's Office	10	11	13	13	13		
(III) Teachers' Common Room	10	11	12	13	14		
(IV) Science Laboratory	11	16	11	12	10		
(V) School Canteen	10	14	10	14	12	14	10
Key:	• Classroom: no trend or characteristic • HOD's Office: recurring data ("13-13-13") • Teachers' Common Room: increasing data • Science laboratory: abnormal datum ("16") • School Canteen: more data points						

The "problem" given in Probe 3(c) was again to compare data from two surfaces (for e.g., HOD's Office and Science Laboratory) and to identify one that showed more pulling force. If the PSTs could not decide, they could claim "I can't tell which one". The experts picked this option since the degrees of variation in all the data sets were too small (and the number of repeats too few) to judge which data set showed more pulling force decisively. As in the previous probe, the mean values for the data sets were all the same, and this again prevented the PSTs from using the value as a routine response.

Table 7.12 shows the results of comparing all pairs of data sets.

Table 7.12 Comparing surfaces for "most" pulling force (N=20)

No.	Number (%)		
(i)	Classroom 4(20)	HOD's Office 5(25)	I can't tell which one 11(55)
(ii)	Classroom 4(20)	Teachers' Common Room 5(25)	I can't tell which one 11(55)
(iii)	Classroom 5(25)	Science Laboratory 4(20)	I can't tell which one 11(55)
(iv)	Classroom 2(10)	School Canteen 3(15)	I can't tell which one 15(75)
(v)	HOD's Office 3(15)	Teachers' Common Room 5(25)	I can't tell which one 12(60)
(vi)	HOD's Office 7(35)	Science Laboratory 5(25)	I can't tell which one 8(40)
(vii)	HOD's Office 2(10)	School Canteen 4(20)	I can't tell which one 14(70)
(viii)	Teachers' Common Room 6(30)	Science Laboratory 4(20)	I can't tell which one 10(50)
(ix)	Teachers' Common Room 4(20)	School Canteen 3(15)	I can't tell which one 13(65)
(x)	Science Laboratory 4(20)	School Canteen 5(25)	I can't tell which one 11(55)

In each comparison, most PSTs picked the option "I can't tell which one" as their choice thus agreeing with the experts' choice. But after the reasons were coded and categorised, the findings showed 42% chose the option mainly because the mean values were the same as compared to only 16% who referred to the variation in the data sets. This affirmed earlier P1 observations and conclusions (see Sections 5.4.3 and 6.4.3).

The data characteristics given in Table 7.11 seemed to have "attracted" only a small number of PSTs in each comparison; so, the numbers were too small to draw any inferences. Thus, instead of looking at each comparison, the significance of a data characteristic can be studied by looking at all the responses together. The coded responses based on a data characteristic (each was used four times) across all the comparisons were summed up to indicate its "attractiveness" and how the PSTs had responded to it (see Table 7.13).

Table 7.13 Analysis of data characteristics

Ideas	Number of coded responses	Quotes from PSTs
Recurring data	18	The last three attempts yielded a constant reading. (Q2-9)
Increasing data	8	Increasing gradient of (III) [Teachers' Common Room]. (Q2-2)
Abnormal datum	2	The value taken at the Science Lab is an anomaly. (Q2-19)
More data points	16	School canteen experiment was conducted more times. (Q2-13)

The two characteristics that seemed to have influenced the PSTs the most were "recurring data" and "more data points". "Recurring data" was not only seen as a *repeating* value (for e.g., "13" in HOD's Office) but also if a value appeared "intermittently" (for e.g., "14" in School Canteen). The PSTs who selected HOD's Office or School Canteen believed they selected a data set that was not only more reliable but the recurring value was also higher than the values in the other set. The notion of "recurring data" seemed to have also influenced several PSTs to choose "I can't tell which one" when one or both data sets contained "irregular" (Q2-8) or "fluctuating" (Q2-10) readings that might be construed as being unreliable.

The PSTs who were influenced by "more data points" would select School Canteen over the other data set. But a large number (10) were not actually swayed by the notion "the more data the better" rather by the repeated appearance of "14" in the data set (thus, again "recurring data"). Interestingly, several PSTs selected only the first five readings of the School Canteen when the data set was being compared since all other data sets contained five repeats; and, some selected "I can't tell which one" because they believed it was "unfair" to compare a pair of data sets with unequal number of readings. There was not much in the PSTs' responses to explain why different number of readings led to "unfairness", but one PST wrote:

> (**V**) [School Canteen] had seven readings while (**I**) [Classroom] had only five which would result in inaccurate comparison of the *average* results. (Q2-6)

Q2-6's idea of "unfairness" came from the notion that the two mean values were unequal in terms of the "quality" of the mean values; she believed the mean value from a larger number of readings would be a better representative. Although the latter has a statistical basis, the idea seemed to have been routinely applied without considering the insignificant difference between the numbers of repeats in the pair of data sets.

Probe 3 was basically set out to see the PSTs' purpose of repeating a reading and how they used certain data characteristics to draw conclusions. Probe 3(a) was able to reveal the understanding that was set out in its objective (see Section 7.4.2). In Probe 3(b), a number of PSTs chose not to select any data set because the mean values of the data sets were the same. On hindsight, It would be interesting to see how these PSTs would have reacted if one of the data sets was set up with larger variation (thus less reliable) but its mean value would be slightly higher than the other. Nevertheless, Probe 3(b) did show the importance of the mean value to the PSTs generally in comparing data sets, and that most PSTs did not look at it in relation to the degree of variation in the data sets.

The problem in Probe 3(b) persisted in 3(c). In retrospect, the ideas in Probe 3(c) could be tested using a different investigative task so as to avoid the participants from being influenced by their earlier responses. Additionally, the number of comparisons, ten in all in Probe 3(c), could be reduced by not testing data characteristics that were already explored in Probe 3(b).

7.5.3 "The Bouncing Rubber Ball Test"

The responses to Probe 4 are reported in Table 7.14.

Table 7.14 Analysis of Probe 4 (N = 20)

Probe 4	Suggested number of readings	About 3	About 10	About 20
		Number (%)		
(a)	**20**	7(35)	11(55)	2(10)
(b)	**3**	13(65)	7(35)	0(0)
(c)	**10**	3(15)	14(70)	3(15)

The results showed only 10% agreed with the experts' option in Probe 4(a), but the percentage increased to 65% and 70% for Probes 4(b) and 4(c) respectively. Based on individual responses, only 2(10%) were in complete agreement with the experts' choices in all situations, and 10(50%) for (b) and (c) only. In contrast, 3(15%) totally disagreed with the experts' choices in all three situations.

The reasons given in Probe 4(a) showed that 35% chose "about 3" readings because they thought taking more than three readings would be futile since the variations were small as the following quote shows:

> The readings do not vary much, and therefore, there is a possibility the repeated values would fall within the same range. (Q2-3)

The two PSTs who selected the suggested option seemed to have the right understanding as they considered the small variation as the reason for the need to take twenty readings in Probe 4(a):

> The variation in their bounces was small ranging only from 25 to 30. Test up to 20 times to observe the differences they can get. (Q2-15)

Table 7.14 shows the option "about 10" was a popular choice but this was only because a group of PSTs consistently picked the "middle" option in all three situations thinking it was the "safest" option:

> Same as answer (a); three repeats do not give us an accurate [reliable] reading. It is unrealistic to conduct 20 repeats. [Q2-19's reason in Probe 4(b)]

In Probe 4(b), the reasons given by those who chose “about 3” (the suggested option) were largely based on being able to see the difference between the balls within the first three measurements:

> By the first 3 readings, [we can tell] ball **D** is bouncier as it re-bounds higher (Q2-2)

Finally for Probe 4(c), amongst those who chose the suggested option of “about 10” readings, only 5(25%) made some reference to variation in the repeats. For instance, Q2-15 felt “ball **F**'s readings varied a lot” and thus required about ten readings to tell the difference; and Q2-20, having described the spread, claimed that only “after the tenth time, we could observe the pattern in the results”.

Probe 4 was designed to explore the PSTs’ ideas on the number of repeats by looking at whether they would look at the mean value in relation to the variation in the repeats. The evidence, however, showed the majority tended to look at the difference in mean values only; they seemed to be able to intuitively suggest the number of repeats when comparing data sets with a large mean value difference in relation to a small degree of variation in the data sets; however, they became less competent as the difference in mean values became smaller. There was a small number of PSTs who completely ignored the mean value or the variation in the data, and suggested a “safe” number of repeats all the time. This reflects their lack of knowledge about the link between the number of repeats and the mean value as well as data spread. Overall, Probe 4 seemed to be effective in meeting its objectives and should be retained in its current format.

7.5.4 "Repeats"

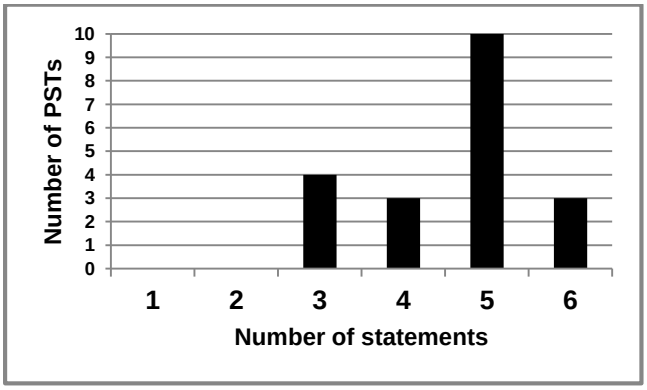
The analysis for Probe 2, "Repeats", is shown in Table 7.15 below.

Table 7.15 Analysis of Probe 2 (N= 20)

Statement		Experts' choice	Agree	Disagree
			Number (%)	
(a)	Three repeated readings are all we need.	**Disagree**	5(25)	**15(75)**
(b)	Most readings when done several times will vary a bit no matter how careful you are.	**Agree**	**20(100)**	0(0)
(c)	We decide the number of repeats after we have done a few readings.	**Agree**	**16(80)**	4(20)
(d)	If you get one reading that is very different from all others you should leave it out of your calculations.	**Agree**	**3(15)**	17(85)
(e)	People who are good at doing experiments always get the same reading each time when making a measurement.	**Disagree**	0(0)	**20(100)**
(f)	The variations in repeated readings are due to human errors only.	**Disagree**	2(10)	**18(90)**

From Table 7.15, with the exception of (d), all other results showed strong agreement with the experts' answers implying the PSTs generally understood the procedural ideas conveyed by these statements. Another analysis was done to determine the level of understanding at the individual level; the data in Figure 7.10 showed a high percentage of 65% agreed with the expert for five out of six statements.

Figure 7.10 Analysis of individual responses (N= 20)



The strong disagreement in (d) was not surprising. Just as in P1Q1, 85% in P2Q2 "extrapolated" the statement and thought a "very different reading" should first be investigated to determine the reason for its large deviation from other readings.

The idea in (d) had been tested twice in the same format and the PSTs still interpreted the statement differently from its actual intention. Perhaps, if the same idea were to be tested again, it could be in a multi-tiered format (see Figure 7.11) to allow the PSTs to state their supporting reason/s for their initial answer:

Figure 7.11 Suggested multi-tier question for 4(d)

(d) If you get one reading that is very different from all others you should leave it out of your calculations.	**Agree**	**Disagree**
Why do you choose that answer in (d)?		

7.5.5 "Starting an Investigation" and "What next in an Investigation"

Table 7.16 shows the analysis of results for Probe 5(a), **"**Starting an Investigation" (see Section 7.4.3).

Table 7.16 Analysis of Probe 5(a) (N = 20)

Data collection sequence	Number (%)
A Trend-focused	5(25)
B Repeats-focused	14(70)
C Repeats-cum-Trend-focused	0(0)
No response	1(5)

The majority (70%) can be classified as "repeats-focused" or "rowers". The reasons given by this group were analysed and the results revealed two categories based on the following claims:

(a) **B** was more efficient and systematic, as implied by the following quote:

> It is more systematic and precise to repeat the experiment immediately than to change it and then to change it back. (Q2-13)

(b) **B** was less susceptible to errors, as implied by the quote below:

> So that minimal variation in the environment can affect a set of [repeated] readings. (Q2-16)

The 25% of PSTs who opted **A** were "trend-focused" (or "columners") and they basically wanted to see if the DV values would change in relation to the IV:

> Since **X** varies, it's better to see how it affects **Y** at different points, and then repeating step (4). (Q2-11)

The results revealed option (**C**), “Repeats-cum-Trend-focused”, was not selected at all, which could mean there were no “arrayers” in the group. Alternatively, the sample size for P2Q2 was just too small to detect an “arrayer”. Overall, the results indicated the majority of PSTs seemed to have no concerns about variation (or uncertainty) in planning their measurements rather they were more focused on being able to complete the measurements efficiently (i.e., by taking repeats after repeats at different intervals, or by taking a sets of DV measurements for a fixed range of IV intervals). The evidence reflects a deeper issue in which the PSTs might not have been well-exposed to planning *open-ended* investigations in their past learning experiences, and were probably quite accustomed to practical procedures that led to guaranteed outcomes. With regards to the structure or format of Probe 5(a), it was effective in revealing the PSTs’ approach in planning measurements for an investigation, and to see whether they were informed by their understandings of uncertainty in measurements.

The next Probe 5(b) would be analysed in two parts. The first, “What next in an Investigation” prompted the PSTs with a table containing a set of “messy” data (see Figure 7.12) accompanied by a question asking the PSTs for their next two readings plus an explanation for their choice. The selection would imply whether the PST was a “columner”, “rower”, or an “arrayer”.

Figure 7.12 Probe 5b (ii)

independent variable X	Dependent variable Y				
	1	2	3	4	5
	F				
20	*90*	*81*	*77*	A	B
30	*73*	*80*	*80*	C	
40	*72*	*79*	*81*		
	D				
	E				

(I) What would be your next **2** readings?

(2) A & B (2) A & C (3) D & E (4) E & F

(II) Why? _____

_____

_____

The results to Probe 5(b) (i) to (iii) are given in Table 7.17 below.

Table 7.17 Analysis of Probes 5(b) (i) to (iii) (N= 20)

5(b)	(1)A&B	(2)A&C	(3)D&E	(4)E&F	No response
(i)	12(60)	5(25)	2(10)	0(0)	1(5)
(ii)	12(60)	3(15)	4(20)	0(0)	1(5)
(iii)	12(60)	5(25)	2(10)	0(0)	1(5)

Several points can be drawn from Table 7.17. First, option (1) "**A&B**" was the most popular choice in all three probes from (i) to (iii). Those who chose "**A&B**" were most likely "repeats-focused" as the two readings were additional data to the data obtained for IV interval "20". Second, the results show there were more "rowers" amongst the PSTs, which was consistent with the previous result shown in Probe 5(a). The reasons given by these "rowers" were also similar to those seen earlier in Probe 5(a) - taking more repeats at a single IV interval would be "systematic":

> It is more systematic and precise. (Q2-13 in 5bi)
>
> I would complete all the readings for one IV first before changing. (Q2-5 in 5bii)

Comparatively, a small group of PSTs in all three probes reasoned they would repeat their measurements only because of the high degree of variation in the DV repeats for the IV interval "20":

> See if the [DV] values continue to vary. (Q2-6 in 5bii)
>
> The [DV] values for **X**= 20 is too wide a range. (Q2-19 in 5biii)

The next popular choice was "**A&C**". These readings were located at the end of the repeats for IV intervals "20" and "30" (see Figure 7.12); therefore, the choice of "**A&C**" might represent either "trend-focused" or "repeats-focused" depending on the explanations (see Figure 7.6). On analysing the latter, all those who chose "**A&C**" seemed to be "repeats-focused". A quote below typifies the pattern of response:

> The DV results seemed to follow a [decreasing] pattern; the next two readings will just confirm this pattern. (Q2-6 in 5biii)

Thus, if the number for "**A&C**" and "**A&B**" in Table 7.17 were to be summed up, the total number of "rowers" should be between 75% and 85%.

The few PSTs who chose "**D&E**" were all "columners", and the quote below exemplifies the pattern of response from this category:

> The first 3 data [in column 1] seemed logical and there was an observable [decreasing] pattern. (Q2-16)

They thought it would only be logical to take readings "**D&E**" in column 1 so that a relationship between the variables could be established. None of the PSTs chose option (4) "**E&F**", which would have identified an "arrayer", one who would probably be concerned with the adequacy of the IV range in establishing a relationship between the variables given the reliability of the repeats were already established.

The next set of two probes 5b(iv) and (v) were more open-ended compared to the previous three as the PSTs were allowed to choose freely their next two readings using any two letters presented in the data table (see Figure 7.13). The accompanying question was also not worded in a way to get the PSTs to interpret what the data was showing (see Annex 3.10). An "arrayer" in these probes could be identified if one letter from the end of the repeats (for e.g., "**F**") and another from the beginning or end of a column of readings (e.g., "**A**" or "**Q**") were picked.

Figure 7.13 Probe 5b (iv)

independent variable X	Dependent variable Y				
	1	2	3	4	5
10	A	B	C	D	E
20	***80***	***81***	***82***	F	G
30	***82***	***89***	***83***	H	I
40	***83***	***84***	***85***	J	K
50	L	M	N	O	P
60	Q	R	S	T	U

(I) What do the data (in numbers) show?

_____

_____

_____

(II) Which letters can represent your next **2** readings and why?

____ &____ because_____

_____

The categorisation of the PSTs was carried out only after the explanations had been carefully analysed. To illustrate, the choice of "**L&M**" in Figure 7.14 could mean both "trend-focused" or "repeats-focused" depending on how the PSTs were to process the tabulated data. For instance, Q2-5 was categorised a "rower" because she said: "I would want to take the [DV] readings for the following IV interval [at 50]", which implied she was merely keen to continue taking the next set of repeats at IV interval "50"; on the other hand, Q2-17 was "trend-focused" because she claimed: "same as (i) [before], the readings suggested the values were recorded *downwards*".

The results of analysing Probe 5b (iv) and (v) are shown below in Table 7.18.

Table 7.18 Analysis of Probe 5b (iv) and (v) (N = 20)

Categories	Numbers (%)	
	5b(iv)	5b(v)
Repeats-focused ("rowers")	14(70)	14(70)
Trend-focused("columners")	4(20)	4(20)
Repeats-cum-Trend-focused("arrayers")	0(0)	0(0)
No response	2(10)	2(10)

The pattern of results seen in Probes 5b (iv) and (v) was similar to those given earlier in Probes 5b (i) to (iii); most PSTs were "rowers", and there were also no "arrayers".

Probe 5(b) was also analysed at the individual level to see if any PST had responded as a "rower" in one probe and a "columner" in another, as such a response could possibly indicate an "arrayer". After analysing all the probes in 5(b), 13(65%) appeared to be consistently "repeats-focused", and 1(5%) was consistently "trend-focused". The six remaining PSTs chose different options across the five probes (see Table 7.19).

Table 7.19 Categorisation of responses to Probe 5(b)

PST	(i)	(ii)	(iii)	(iv)	(v)
Q2-1	R	T	R	R	R
Q2-2	R	T	R	R	R
Q2-10	R	R	R	T	T
Q2-11	R	R	R	R	T
Q2-16	R	R	R	T	R
Q2-20	R	R	R	T	T

Key: **R** = “Repeats-focused”; **T** = “Trend-focused”

Just by numbers alone, Table 7.19 shows there is a high probability that most of the PSTs could be “rowers”. Some PSTs (for e.g., Q2-10 and Q2-20) could be quite sophisticated in processing “messy” data as they switched between different focus for different sets of data. These individuals could potentially be “arrayers” but more evidence (for e.g., testing with a larger number of “messy” data sets) might be necessary to reach a conclusion.

Finally, for Probe 5(b), there was relatively a higher number of “nil” response and “signs” (for e.g., question marks, cancellations, ditto, etc.) that seemed to indicate the PSTs had difficulties processing the “messy” data. On hindsight, the difficulties could be mitigated by modifying the data in the tables to a single from a double digit to make it easier for the PSTs to process the data.

7.6 Summary of Chapter

Chapter 7 mainly focused on addressing Research Aim 2, which was about the early steps in developing a questionnaire that would allow the accurate interpretation of the PSTs’ understanding of uncertainty in measurements. It described Phase 2 of this research where findings from the earlier Phase 1 studies were drawn and used in the design and development of probes in the questionnaire. The chapter also described the results and findings of testing the questionnaire with the intent of analysing the probes for efficacy and making improvements.

CHAPTER 8

CONCLUSIONS

8.1 Introduction

The study sets out to explore and describe PSTs' understanding of uncertainty in measurements taken during science investigations, and to develop a questionnaire that could be used to identify patterns and divergences in the PSTs' understanding of the concept. In order to do these, the study first identified the measurements the PSTs would likely have to take based on the types of variables found in an investigation, and then adopted the Concepts of Evidence (Gott & Duggan, 1995; Gott et al., 2008) as a theoretical framework for developing and analysing probes that sought evidence of procedural ideas underlying the understanding of uncertainty in measurements. The research process followed the guiding principles proposed by Johnson and Gott (1996): triangulation of evidence; creation of "neutral" probes; and analysing the response data using the participants' frame of reference with the purpose of developing an accurate interpretation of the PSTs' procedural ideas. In order to develop the research method to answer its research questions (including the development of a questionnaire), the study carefully studied the literature to look at limitations other studies might have and avoided them.

Besides contributing to the general literature on uncertainty in measurements, the research specifically filled the "gaps" in knowledge largely in two areas. First, the study was on *pre-service primary teachers* and about exploring their procedural ideas underpinning uncertainty in measurements; perhaps, there might be few studies bearing such intention conducted on these subjects, but none to my knowledge could be found in the local research academia. Second, the study looked beyond understanding uncertainty in

single measurements or repeated measurements responding to a categoric IV which were routinely being investigated in other studies, and explored understanding of uncertainty in repeated DV data responding to a range of IV intervals.

In the following discussions, I shall conclude by outlining several key points about the research aims. I shall also discuss the implications this research holds for theory and for practice. The limitations that bounded the impact of this study will also be highlighted together with some recommendations for future research.

8.2 Conclusions about Research Aims

Research Aim 1 was concerned with exploring and describing the PSTs' understanding of uncertainty in measurement whereas Research Aim 2 looked towards the development of a questionnaire that would show the patterns and divergences in the PSTs' understanding of the concept. Five research questions were crafted to address these aims; the following discussions shall look at how each has been answered.

(a) The first intended to explore the PSTs' understanding of the inherent variability of measurements. The results from P1Q1 showed 87% believed "it is never possible to repeat a measurement or reading in exactly the same way". Additionally, about 95% from the same study and 100% in P2 expected to see variation in repeated readings. The same pattern of thinking seen in these results was reinforced by evidence from P1I1 and P1I2 where almost all PSTs indicated variation in their repeated data for the "Bouncing Ball" investigation.

Other evidence came from questions that looked into the belief in true values. Only about 18% in P1I1 claimed such values existed, thereby discounting uncertainty in measurements, but this also meant the majority

expected uncertainty in all measurements. An almost similar percentage in P1I1 and P1I2 believed a “perfect” measuring instrument could actually exist. The P1I1 results showed the PSTs who held such a notion would most likely believed that true values existed as “real” values.

The understanding of uncertainty was premised on being able to apply key underpinning concepts: accuracy and precision. However, there was confusion between the two; many PSTs used “accurate” instead of “precise” to refer to close readings (which they described as “consistent data”) when they *applied* their procedural ideas to describe the purpose of repeated measurements in the “Bouncing ball” investigation (P1I1 and P1I2). However, they responded correctly to questionnaire statements in P1Q1 that looked at the conceptual definition of precision or were able to recall the conceptual meaning of accuracy in P1I2. This could imply the understanding of these concepts was only at the recall level, and not higher. Besides, the majority of PSTs did not seem to understand the interrelationships between the two fundamental concepts as investigated in P1I2.

(b) On selecting a measuring instrument to take a single measurement, the PSTs were able to intuitively select instruments that gave the most accurate measurements. In choosing static-scale instruments, the PSTs used the idea of “resolution of scale” more than “full-scale deflection” or “percentage error”. This was constantly shown across different probes in P1Q1, P1I2 and P2Q2. In P2Q2, the PSTs also showed understanding about “limits of detection” and “scale readability”, ideas used for selecting digital instrument like a weighing balance. Another probe, however, showed some PSTs selected digital over analogue clocks simply because of its sophisticated appearance, which they

misconceived as implying better accuracy although both were equivalent in telling the time right down to the seconds.

(c) The purpose of repeated measurements was probed several times in P1, and only a small number of PSTs was consistently found to have understood the purpose of repeats was to “capture” variation in data. The majority, however, suggested they were either looking for consistent data or to get readings for a mean value calculation. For most PSTs who sought consistent data, they understood that such data represented reliability, but a few would claim they were actually looking for a recurring value that represented the actual measurement, in other words, the true value.

The causes of variation was the subject of several probes in both P1 and P2; the evidence showed most PSTs generally attributed the causes of variation (in descending order) to random human errors, uncontrolled variables, and the inherent characteristics of the measuring instruments. For a few P1 PSTs, human errors were the only cause of variation. In P1I1 and P1I2, there was evidence of a handful of PSTs whose idea of human errors might be flawed since they thought of them as “mistakes” that could be eliminated by using a better technique of measurement.

The number of repeated measurements was another key issue. The notion of “the more, the better” was quite widespread but the evidence showed the PSTs did not know about the supporting reasons such as standard error. Statistical concepts (other than the mean value) like standard deviation and normal distribution were hardly mentioned. The number of repeats that should be decided by the degree of variation was instead determined by logistical considerations like time and manpower requirements or by routine practices of a fixed number of repeats (as high as 25% in P1I1 and P2Q2 had such an

idea). When the PSTs were asked to compare data sets, very few looked at the difference in mean values of the data sets in relation to their degree of variation to determine the number of repeats. The majority (in P2Q2) seemed to be using only the mean value and did not consider variation in data at all. Thus, when the mean value difference in the probes became smaller or even zero, the PSTs became much less competent in distinguishing the sets. For several PSTs in P1I2, P1Q1 and P2Q2, they also felt it was "unfair" to compare data sets with unequal number of repeats (also observed in Lubben & Millar, 1996), and as a result, some might not consider the supposedly "extraneous" data.

The idea of reliable data was also examined by looking at the PSTs' selection of data based on their characteristics. The P2Q2 results showed most PSTs tended to select a data set with a recurring value as the most reliable and "trustworthy" (which explained why they normally select such data in other probes when given a choice).

Finally, the idea of "abnormal" data was studied in P1Q2 and P2Q2. Both found the PSTs indicating they would like to check the cause of the large deviation before deciding what to do with the data. However, there was no mention of keeping the data if it was identified as part of the variation, thus reflecting the idea was rather uncommon.

(d) Three categories were derived based on the way the PSTs processed "messy" data (counterintuitive with high degree of variation). The largest category belonged to the "rowers" ("repeats-focused") who were inclined on getting more repeats so that precise readings could be obtained to calculate a mean. The evidence in P1I2 and P2Q2 indicated the "rowers" would likely plan to complete a set of repeats for one IV interval before moving on to the next. The second category, the "columners" ("trend-focused"), tended to look up or

down the DV measurements taken for the whole range of IV intervals to find a relationship between the variables. The "columners" would likely plan their data collection by taking a single measurement for each IV interval starting from the smallest; and repeated measurements were likely meant for replacing data that did not fit the trend. A third category known as the "arrayers" ("repeats-cum-trend-focused") might exist but the studies did not reveal any. This group would be informed by their understanding of uncertainty and their measurement plan would likely consider establishing the IV range and the degree of variation by conducting "preliminary trials". The fact that "rowers" and "columners" existed implied the PSTs were not accustomed to carrying out "preliminary trials" to plan their measurements in their past learning experiences. Additionally, the PSTs might not be well-exposed to planning an investigation, and were likely to have been given experiments where the outcomes were already known.

(e) The final part throws light on the research problem that motivated this research as it focused on the PSTs' ability to articulate their understandings of uncertainty in measurements through the questionnaire. There was evidence to suggest the majority did not have a full range of procedural ideas to handle uncertainty in measurements. Certain concepts like "standard error" and "percentage error" were absent whereas others like "variation" and "full-scale deflection did not seem to have been fully developed as only a few PSTs were able to explain these concepts in different situations. On handling uncertainties in measurements, the PSTs largely relied on routine ideas without much understanding about the thinking behind those ideas. A good example would be the "more measurements, the better" that many seemed to know but did not understand what "better" really meant (also observed in Séré et al., 1993). Another example was "mean value", which one PST in P1I1 described as a

"cultural practice". A third example was "experimental error", which another PST claimed should always be present because school-based investigations would always asked why it was present and not why it was absent. Finally, the "rowers" and "columners". The rather fixed pattern of processing data might have come about as a result of constantly being trained in a particular way of collecting data during investigations.

8.3 Implications of study

8.3.1 Understanding uncertainty in measurements

Several findings in the literature need to be revisited in the light of evidence provided by this study. Several researchers believed some students might be seeking true value(s) when they repeated their measurements (Allie et al., 1998; Coelho & Séré, 1998) to look for "consistent data". The South African studies categorised such students as "point" reasoners. In this study, a significant number of PSTs were also observed to be seeking "consistent data", but what could be construed as seeking for true value(s) was actually a strong effort towards getting *very* "precise" readings. The PSTs were in fact very outcome-driven, setting a high target of recurring data as their goal, but might be contented if the readings appeared very close. Thus, in the light of this observation, this study agrees with the findings of Heinicke and Reiss (2001) that found their subjects mostly repeated their measurements to check previous results in order to see if the data were consistent.

Lubben and Millar (1996) proposed the selection of recurring data should be seen as an early step in the "progression" of choosing repeated data that are closed together. The latter was seen as an advanced step because it demonstrated the understanding of precision. In this study (see P1I2), the results showed the selection of a recurring reading ("consistent data" in the

words of the PSTs,) was not a nascent stage; rather, the PSTs viewed recurring reading as a step higher than close readings and indicated the measurement process was “good enough” for high quality data.

The South African studies (Allie et al., 1998; Buffler et al., 2001) proposed the concepts of “range” and “mean value” to indicate “spread thinking” and characterised a “set reasoner”. The researchers, however, acknowledged the difficulty in using the criteria to distinguish students as “set reasoners” as the evidence showed both concepts were being routinely applied without real understanding. This study agrees with the observations but would argue the understanding of uncertainty especially for tertiary students including PSTs should include the idea of randomness, and that repeated measurements were normally distributed and subjected to the laws of probability. Tertiary students should have a deeper understanding of the concept of variation along with underlying statistical concepts (for e.g. SD and SE) that support the handling of uncertainty in repeated measurements. This implies the testing of such concepts should be included in any instrument that evaluates tertiary students’ understanding of uncertainty.

In selecting values of the IV using pendulum and forcemeter investigations, Kanari and Millar (2004) suggested 85% of their subjects were mainly “trend-focused”, and to a smaller extent, “difference-focused” (trialling of a pair of high and low IV intervals to see if this resulted in a difference in the value of the DV). It would be interesting to see how the participants would respond if investigative tasks were totally unfamiliar to them. This study using “messy” data showed the majority of PSTs were “repeats-focused” (“rowers”) instead of “trend-focused”, but when “messy” data from a familiar investigation like the pendulum experiment were used, the PSTs became more “trend-

focused" because of the influence of prior knowledge. This should be an important consideration in the development of probes, which is further discussed in the next section.

8.3.2 The nature of probes in assessing understanding

In this study, a range of probes were sometimes used to assess the same procedural idea, two important aspects seemed to have impacted the PSTs' ability to respond. One was the context of the probe. In P1I1, for instance, the use of a biological context heightened the PSTs' awareness of uncertainty, and might have even prompted the PSTs to claim the live specimen (apples) instead of "human errors" (which were often cited) was the major cause of variation.

This study also agrees with the findings from Leach et al. (1998) in that it observed the PSTs were more able to articulate their procedural ideas and apply them in investigative task-based probes compared to generalised-based ones (also known as " decontextualized probes"). For instance in P1I2, many PSTs had difficulty articulating the conceptual meaning of precision when they were asked directly. Besides, they often struggled if they chose to respond with an example. However, in the same study, they were able to illustrate precision and described their procedural ideas when they responded in the "Bouncing ball" investigation. This implied the PSTs' might also have some tacit understanding of the concept that they found easier to apply to a situation.

The second aspect was the purpose of measurement. The evidence from this study concurred with Lubben et al. (2004) that found students might base their judgements about the quality of data on the perceived purpose of the measurement. This could be seen in the use of a digital weighing balance to weigh a fish in P2Q2 where some PSTs assumed that an everyday situation (for

e.g. the market) should demand less accuracy; but, if the measurement was meant for a laboratory investigation, a higher accuracy could be expected. The same issue emerged when the PSTs had to choose between an analogue and a digital clock.

8.3.3 Teaching procedural ideas

To some extent, the study showed the PSTs generally did not have *deep* conceptual understanding about uncertainty in measurements; there was a lack of "integrated or holistic understanding" (Garfield & Ben-Zvi, 2005) which would have allowed the PSTs to discover relationships, solve new problems, construct explanations, and draw conclusions. The evidence also showed the PSTs were mainly reciting "fragmented" ideas and following learned procedures in familiar situations rather than applying and adapting an idea to new probes, or one with a different context. This has implications for studies that intend to use a single context to evaluate understandings of uncertainty in measurements. But more critical to this study are the implications for the way uncertainty in measurements was taught to students (including the PSTs in this study).

The science curriculum in Singapore adopts a skills-based approach, which is characterised by performance often termed as "process skills", to teach practical skills (MOE, 2008). The main assumption of this approach is that procedural understanding could be learned and acquired through practice with lots of experiments. The teaching of "procedural ideas" is largely implicit in the teaching of substantive concepts, and any guidance rendered to students is only through simple exemplification of the process. Investigations are usually more outcome-driven and mostly meant to illustrate substantive concepts; teachers therefore tended to downplay the procedural component, giving

specific instructions in order for the experiments to lead to the substantive ideas. There is very little chance that the students would get to deal with design decisions or interpreting “messy” data (Roberts & Gott, 2002). Many researchers, however, have shown that explicit teaching of ideas which contribute to procedural understanding (including the uncertainty of measurements) seems to be effective and does develop understanding (Gott & Duggan, 1995, 2003; Roberts & Gott, 2004; Roberts & Gott, 2006; Glaesser et al., 2009a, b). Such teaching can either involve science investigations in the laboratory or through didactic teaching that include the use of data probes. For PSTs in particular, courses that explicitly teach procedural understanding can be mounted so that they are able to learn to teach their students procedural ideas in the future.

Based on the findings from this study, a brief outline of a programme for teaching ideas of evidence related to uncertainty in measurements is proposed (Annex 8.1) for the Singapore context. The ideas for this programme are substantially borrowed from the “Evidence Module” designed for the BAEd Course at Durham University (Gott et al., 2008). Although the teaching programme is targeted at PSTs, it can be modified for in-service primary teachers. Two basic conditions, however, must be satisfied before such a programme can take place, and these are: (a) the participants must have sufficient substantive understanding of scientific concepts (at least those in the primary science syllabus); and, (b) they must have some knowledge of what a scientific investigation is; and, are able to identify the IV, DV, and the CV(s) in an investigation as well as to record the IV and the DV values in a table format. To my knowledge, given the current teacher preparatory programme in the National Institute of Education (Singapore) and the stringent selection criteria

for entry into the programme, such conditions should be easily met. Besides, all teachers would have good exposure to working in the laboratory (albeit dealing mostly with recipe-like practical work) in their past learning experiences (see Sections 1.3 and 3.3).

8.4 Limitations of study

- One of the limitations of this study was not all areas of procedural understand were covered in the instruments (for e.g. the understanding of graphs and charts), the conclusions, therefore, referred specifically only to those that were examined in the study. The focus on some areas and not others was deliberate so that more depth could be given to the discussions on the chosen areas within the constraints of time and word limit. Besides, there was always this notion the areas not covered could be part of a future research.
- Some may argue that the sample of fifty-five PSTs in P1 and twenty PSTs in P2 in this study may be too small to represent the entire population of pre-service primary teachers, and to generalise its findings. However, the results and implications of this research can be relevant and meaningful if one adopts the view based on the concept of relatability (see Bassey, 1981) in which the PSTs in this research could be seen to be like any other pre-service teachers present in other teacher training institutions in other parts of the world. Bearing such a perspective, the academics or policy-makers might therefore be able to recognise their own PSTs amongst the participants of this research, and thereby relate the findings to their own settings.
- The goals of this study would have been better met if the questionnaire was passed through more rounds of testing and refinements but time

did not allow for this. In addition, I was not able to carry out follow-up interviews with the P2 participants to validate certain findings and make improvements to Questionnaire 2.

- The objective of achieving an accurate interpretation of the PSTs' responses to the instruments was crucial to this study. Although the methodology, among its many goals, was geared towards refining the researcher's interpretation, all the PSTs' responses were interpreted alone by me. Despite regular checks with supervisors and colleagues on the interpretation of responses, a more systematic way of checking for reliability of interpretation such as finding out the level of agreement between the researcher and other parties could have helped to address the issue to some extent.
- Finally, the conclusions for this research were drawn based on evidence obtained via interviews and questionnaires. Lubben and Millar (1996) cautioned that some aspects of procedural understanding such as the significance of small variance in repeated measurements could be better revealed with laboratory investigations. I agree with this view and believe the use of laboratory could have given more of the same evidence that was revealed in this study, and perhaps, some finer points as those mentioned by Lubben and Millar (1996).

8.5 Recommendations for further research

- A particular area of procedural understanding that was not covered in this study was on data presentation using graphs and bar charts. I suspect the understanding of uncertainty might have an impact on the way the PSTs would process and select values for their graphs or charts (presumably from a set of values) or even plotting them. Several

studies have indicated this. For instance, the study by Lubben et al. (2001) had shown that students might join all the points on a graph by multiple line segments or draw a single line through “trusted” points. The research literature had also shown that students might have difficulties interpreting charts to distinguish “real” changes in DV measurements due to changes in the IV only (see Millar, 1999; Gott & Duggan, 2003).

- The analyses of the instruments were geared towards simplifying the evidence for the whole sample to a few categories of understandings so that they could be used to guide the design of the probes or the multiple options offered in them. With a more refined P2Q2 questionnaire (based on suggestions given in Chapter 7), it could be used to describe individual PST’s procedural understanding across different areas, and a profile of the individual in terms of understanding of uncertainty in measurements could be drawn to see the patterns of ideas. A follow-up interview with identified individual based on the results could then explore potential causal factors, for example, the effects of epistemic views of the nature of measurements and the use of evidence to explain theory (see for e.g., Leach, 1999; Ryder & Leach, 1999; Ryder & Leach, 2000). Of particular interest to this researcher also would also be factors that contributed to the PSTs being “rowers”, “columners”, and “arrayers”.
- This researcher sees a potential in using the refined P2Q2 questionnaire on other subjects, in particular, upper primary and secondary students, as well as in-service primary teachers. This would be possible for students as the design of the instrument considered

several factors including the use of simple language and the minimal substantive knowledge required to respond to the probes. If the questionnaire is applied for cross-age studies, the results can be used to assess students' progression in the understanding of uncertainty in measurements or the impact on students' understanding in the concept as a result of introducing into the curriculum inquiry science that uses scientific investigation as its main activity. As for primary teachers, their responses to the questionnaire can be used to assess their preparedness in using measurements in their inquiry-based instructions, which served to inform my current role as a Master Teacher.

REFERENCES

Abd-El-Khalick, F., BouJaoude, S., Duschl, R. A., Hofstein, A., Lederman, N. G., Mamlok, R., Niaz, M., Treagust, D., Tuan, H. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397-419.

Abd-El-Khalick, F., Bell, R.L., & Lederman, N.G. (1998). The nature of science and instructional practice: Making the unnatural natural. *Science Education, 82*(4), 417–437.

Abruscato, J., & DeRosa, D. A. (2010). *Teaching Children Science. A Discovery Approach.* Boston, MA: Allyn & Bacon.

Åkerlind, G., McKenzie, J., & Lupton, M. (2011). *A threshold concept focus to curriculum design:supporting student learning through applicationof variation theory.* Sydney, AUS: Australian Learning and Teaching Council.

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

American Association for the Advancement of Science (AAAS) (1993). *Benchmarks for science literacy.* New York, US: Oxford University Press.

American Weigh Scale. (2011). *Weighing Scale Terminology*. Retrieved December 05, 2014 from http://www.awscales.com/support/terminology.

Arksey, H., & Knight, P. (1999). *Interviewing for social scientists: An introductory resource with examples.* London,UK: SAGE.

Australian Curriculum, Assessment and Reporting Authority (ACARA) (2013). *Proficiency Levels - Science Literacy*. Retrieved August 28, 2013, from http://www.nap.edu.au/nap-sample-assessments/about-each-domain/science-literacy/napsa-proficiency-levels---science-literacy.html.

Bassey, M. (1981). Pedagogic Research: on the relative merits of search for generalisation and study of single events. *Oxford Review of Education*, *7*(1), 73–94.

Bell, S. (1999). *Measurement. Good Practice Guide No. 11(Issue 2).A Beginner's Guide to Uncertainty of Measurement*. Retrieved May 15, 2013 from http://www.wmo.int/pages/prog/gcos/documents/gruanmanuals/UK_NPL/mgpg11.pdf.

Buffler, A., Allie, S., Lubben, F., & Campbell, B. (2001). The development of first year physics students' ideas of measurements in terms of the point and set paradigms. *Internationation 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 measurment in the first- year physics laboratory. *Conference of the European Science Education Research Association.* Noordwijkerhout, The Netherlands.

Bybee, R. W., Powell, J. c., & Trowbridge, L. W. (2008). *Teaching Secondary School Science. Strategies for Developing Scientific Literacy.* Upper Saddle: Pearson Education Inc.

Campbell, B., Buffler, A., Lubben, F., & & Allie, S. (2005). *Teaching scientific measurement at university:understanding student's ideas and laboratory curriculum reform. A monograph of the African Journal of Research in Mathematics, Science and Technology Education.* Pretoria,SA: Southern African Association for Research in Mathematics, Science and Technology Education (SAARMSTE).

Chin, C., & Kayalviszhi, G. (2002). Open-ended investigation in Science: A case study of primary 6 pupils. *Journal of Science and Mathematics Education in South-East Asia*, *25*(1),70-94.

Clackson, S. G., & Wright, D. K. (1992). An appraisal of practical work in science educagtion. *School Science Review, 74*(266), 39-42.

Coelho, S., & Séré, M. (1998). Pupils' Reasoning and Practice during Hands-on Activities in the Measurement Phase. *Research in Science & Technological Education, 16*(1), 79-96.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data. Complementary research strategies.* London, UK: SAGE.

Collins English Dictionary (2014). *Complete & Unabridged (10th Edition).* Retrieved February 07, 2014, from http://dictionary.reference.com/browse/approximate

Creswell, J. (2008). *Educational Research. Planning, conducting, and evaluating quantitiative and qualitative research.* Upper Saddle River,NJ: Pearson Education.

Creswell, J. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* Thousand Oaks,CA: SAGE.

Denzin, N. K. (1978). The research act: A theoretical introduction to sociological methods. New York, NY: McGraw-Hill.

Department of Education and Skills, U. K. (DES) (2013). *Science Key Stage 1: Sc 1 Scientific Enquiry*. Retrieved October 20, 2013 from: http://www.education.gov.uk/schools/teachingandlearning/curriculum/primary/b00199179/science/attainment.

Drever, E. (1995). *Using semi-structured interviews in small scale research: a teachers' guide.* Edinburgh,UK: SCRE.

Duerdoff, I. (2009). Teaching uncertainties. *Physics Education, 44*(2), 138-144.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking Science to School. Learning and Teaching Science in Grades K-8. A Report from Committee on Science Learning, Kindergarten Through Eight Grade.* Washington, DC: The National Academies Press.

Evangelinos, D., Psillos, D., & Valassiades, O. (2002). An investigation of teaching and learning about measurement data and their treatment in the introductory Physics laboratory. In D. Psillos, & H. Nieddere (Eds.), *Teaching and Learning in the Science Laboratory* (pp. 179 – 190). Dordrecht: Kluwer Academic Pub.

Evangelinos, D., Valassiades, O., & Psillos, D. (1999). Undergraduate students' views about approximate nature of measurement results. In M. Komorek, H. Behrendt, R. Dahncke, R. Duir, W. Graber, & A. Kross (Eds.), *Proceedings of the Second International Conference of ESERA* (pp. 208-210). Kirl: IPN.

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

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal, 4*(1), 92-99.

Geertz, C. (1973). Thick Description: Toward an Interpretive Theory of Culture. In *The Interpretation of Cultures: Selected Essays* (pp. 3-30). New York,NY: Basic Books.

Glaesser, J., Gott, R., Roberts, R., & & Cooper, B. (2009a). The roles of substantive and procedural understanding in open-ended science investigations: Using fuzzy set qualitative comparative analysis to compare two different tasks. *Research in Science Eucation, 39*, 595–624.

Glaesser, J., Gott, R., Roberts, R., & Cooper, B. (2009b). Underlying success in open-ended investigations in science: using qualitative comparative analysis to identify necessary and sufficient conditions. *Research in Science & Technological Education, 27*(1), 5-30.

Gott, R., & Duggan, S. (1995). *Invetigative Work in the Science Curriculum.* Buckingham,UK: Open University Press.

Gott, R., & Duggan, S. (2003). *Understanding and Using Scientific Evidence. How to Critically Evaluate Data.* London,UK: SAGE.

Gott, R., Duggan, S.,& Roberts, R. (2008). *Concepts of evidence and their role in open-ended practical investigations and scientific literacy; background to published papers.* Retrieved on Jan 20, 2010 from: http://www.dur.ac.uk /rosalyn.roberts/Evidence/CofEv_Gott%20et%20al.pdf.

Gott, R., Duggan, S., Roberts, R., & Hussain, A. (2014). *Research into Understanding Scientific Evidence*. Retrieved on Jan 11, 2015 from: http://community.dur.ac.uk/rosalyn.roberts/Evidence/CofEv_Gott%20et%20al.pdf.

Gott, R., Foulds, K., Johnson, P., Jones, M., & Roberts, R. (1997). *Science Investigations 1.* London,UK: Collins Educational.

Gott, R., Foulds, K., Roberts, R., Jones, M., & Johnson, P. (1999). *Science Investigations 3.* London,UK: Collins Educational.

Guare, C. (1991). Error, Precision, and Uncertainty. *Journal of Chemical Education, 68*(8), 649-652.

Guerra-Ramos, T., Ryder J., & Leach, J. (2010). Ideas about the nature of science in pedagogically relevant contexts: insights from a situated perspective of primary teachers. *Science Education*, 94(2), 282-307.

Heinicke, S., & Heering, P. (2013). Discovering Randomness, Recovering Expertise: The Different Approaches to the Quality in Measurement. *Science & Education, 22*, 483–503.

Heinicke, S., & Riess, F. (2011). Missing links in Experimental Work: Students action and reasoning in measurement and uncertainty. In C. Bruguie`re, & D. Berger (Eds.), *European Science Education Research Association (ESERA).*

Heisawn, J., Songer, N. B., & Lee, S. (2007). Evidentiary Competence: Sixth Graders' Understanding for Gathering and Interpreting Evidence in Scientific Investigations. *Research in Science Education, 37*(1), 75–97.

Hogan, D., Luke, A., Kramer-Dahl, A., Lau, S., Liau, A., & Koh, K. (2006). *Core research program:Year Two Progress Report.Unpublished Curriculum Review and Policy Planning (CRPP) Technical Report.* National Institute of Education (Singapore): Nanyang Technological University.

Hogan, K., & Maglienti, M. (2001). Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions. *Journal of Research in Science Teaching, 38*(6), 663–687.

House of Commons, Science and Technology Committee (2002). *Science education from 14 to 19. Third report of session 2001-2 (1)*. London, UK: The Stationery Office.

Jarvis, T., Pell, A., & McKeon, F. (2003). Changes in Primary Teachers' Science Knowledge and Understanding during a Two Year In-service Programme. *Research in Science & Technological Education, 21*(1), 17 - 42.

Jick, T. (1983). Mixing Qualitative and Quantitative Methods: Triangulation in Action. In J. Maanen (Ed.), *Qualitative Methodology.* Newbury Park, CA: SAGE.

Johnson, P. (2013). Scientific Enquiry (Investigations). *Unpublished Lecture Notes PGCE.* Durham University (United Kingdom).

Johnson, P., & Gott, R. (1996). Constructivism and Evidence from Children's Ideas. *Science Education, 80*(5), 561-577.

Joint Committee for Guides in Metrology (JCGM) (2008). *Evaluation of measurement data - Guide to the expression of uncertainty in measurement.* Retrieved on Sep 24, 2012 from: http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf.

Joint Committee for Guides in Metrology (JCGM) (2012). *International vocabulary of metrology – Basic and general concepts and associated terms (VIM). 2008 Version with Minor corrections* (3rd Edition). Retrieved on Jan 15, 2013 from: http://www.bipm.org/utils/common /documents/jcgm/JCGM_200_2012.pdf.

Jones, M., & Gott, R. (1998). Cognitive acceleration through science education: alternative perspectives. *International Journal of Science Education, 20*(7), 755-768.

Kanari, Z., & Millar, R. (2004). Reasoning from Data: How Students Collect and Interpret Data in Science Investigations. *Journal of research in Science Teaching, 41*(7), 748-769.

Kim, M., Tan, A.L., & Talaue, F.T. (2013). New Vision and Challenges in Inquiry-Based Curriculum Change in Singapore. *International Journal of Science Education, 35*(2), 289-311.

Kirkup, L., & Frenkel, R. (2006). *An Introduction to Uncertainty in Measurment Using the GUM (Guide to the Expression of Uncertainty in Measurement).* New York, NY: Cambridge University Press.

Kirschner, P., & Meester, M. (1988). The Laboratory in Higher Science Education. *Higher Education, 17*, 81-98.

Krathwohl, D. (2002). A Revision of Bloom's Taxonomy:an overview. *Theory into Practice, 41*(4), 212-218.

Kung, R., & Linder, C. (2006). University students' ideas about data processing and data comparison in a physics laboratory course. *Nordic Studies in Science Education (NorDiNa), 2*(2), 40-53. Retrieved December 2, 2012 from: https://www.journals.uio.no/index.php/nordina/article/view/423/485.

Kvale, S. (2008). *Doing Interviews.* Los Angeles,CA: SAGE.

Laugksch, R. C. (2000). Scientific Literacy: A Conceptual Overview. *Science Education, 84*(1), 71-94.

Leach, J. (1999). Students' understanding of the co-ordination of theory and evidence in science. *International Journal of Science Education, 21*(9), 789-806.

Leach, J. (2002). The Use of secondary Data in Teaching about Data Analysis in a First Year Undergraduate Biochemistry Course. In D. Psillos, & H. Nieddere (Eds.), *Teaching and Learning in the Science Laboratory* (pp. 179 – 190). Dordrecht: Kluwer Academic Pub.

Leach, J., Millar, R., Ryder, J., Sere, M., Hammelev, D., Niedderer, H., & Tselfes, V. (1998). Students' images of science as they relate to labwork learning. *Labwork in Science Education, Working Paper 4, Targeted Socio-Economic Research Programme, Project PL 95-2005*.

Lee, K. W., Li, T. L., Goh, N. K., Chia, L. S., & Chin, C. (2002). Science Teachers and Problem Solving in Primary Schools. In A. G. Tan, K. W. Lee, N. K. Goh, & L. S. Chia (Eds.), *New Paradigms for Science Education. A Perspective of teaching Problem Solving, Creative teaching and Primary Science Education* (pp. 192-207). Singapore: Prentice-Hall.

Lin, H. S., Chiu, H. L., & Chou, C. Y. (2004). Student understanding of the nature of science and their problem-solving strategies. *International Journal of Science Education, 26*(1), 101–112.

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

Lubben, F., Campbell, B., Buffler, A., & Allie, S. (2004). The Influence of context on judgements of the quality of experimental measurements. In A. Buffler, & R. Laugksch (Ed.), *Proceedings of the 12th Annual Conference of theSouthern African Association for Research in Mathematics, Science and Technology Education* (pp. 569-577). Cape Town: SAARMSTE.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education, 18*(8), 955-968.

Lythcott, J., & Duschl, R. (1990). Qualitative research: From methods to conclusions. *Science Education*, *74* (4), 445 -460.

Masnick, A., & Morris, B. (2008). Investigating the Development of Data Evaluation: The Role of Data Characteristics. *Child Development, 79*(4), 1032–1048.

McClelland, J. A. G. (1984). Alternative frameworks: Interpretation of evidence. *European Journal of Science Education*, *6*(1), 1-6.

McComas, W., & Olson, J. (1998). The nature of science in international science education standards documents. In W. McComas (Ed.), *The nature of science in science education: Rationales and strategies (pp. 41-52)*. Dordrecht: Kluwer.

Meyer, J.H.F. and Land, R. (2005) Threshold concepts and troublesome knowledge (2): Epistemological considerations and a conceptual framework for teaching and learning. *Higher Education*, 49, 373-388.

Miles M. B., & Huberman, A. M. (2002). *The qualitative researcher's companion*. Thousand Oaks, CA: SAGE.

Millar, R. (1998). Students' understanding of the procedures of scientific enquiry. In A. Tiberghien, E. Jossem, & J. Barojas (Eds.), *Connecting Research in Physics Education with Teacher Education. An I.C.P.E. Book.* The International Comission of Physics Education.

Millar, R. (1999). Understanding how to deal with experimental uncertainty: a 'missing link' in our model of scientific reasoning? *Paper presented at the conference of the European Science Education Research Association (ESERA)*, Kiel, Germany, 31 August – 4 September.

Millar, R., Lubben, F., Gott, R., & 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.

Millar, R., & Osborne, J.F. (Eds.) (1998). *Beyond 2000: Science Education for the Future*. London: King's College London.

Ministry of Singapore (MOE) (2001). *Science Syllabus (Primary)*. Singapore: Sciences Branch, Curriculum Planning & Development Division.

Ministry of Singapore (MOE) (2008). *Science Syllabus (Primary)*. Singapore: Sciences Branch, Curriculum Planning & Development Division.

Ministry of Education (MOE) (2009). A *Guide to Teaching and Learning of Primary Science*. Singapore: Curriculum and Planning Division.

Munier, V., Merle, H., & Brehelin, D. (2012). Teaching Scientific Measurement and Uncertainty in Elementary School. *International Journal of Science Education, iFirst Article*, 1–32.

Murcia, K., & Schibeci, R. (1999). Primary student teachers' conceptions of the nature of science. *International Journal of Science Education, 21*(11), 1123–1140.

National Research Council (NRC). (2000). *Inquiry and the National Science Education Standards. A guide for teaching and learning.* Washington, DC: National Academy Press.

Newton, D.P. (2000). What do we mean by teaching for understanding? In L.D. Newton (Ed.), *Meeting the standards in primary science. A guide to ITT NC*. London: Routledge-Falmer.

Newton, L.D. (2001). Teaching for Understanding in Primary Science. *Evaluation & Research in Education*, 15(3), 143-153.

Organisation for Economic Co-operation and Development (OECD) (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing. Retrieved 15 May, 2015 from http://www.oecd-ilibrary.org/education/pisa-2012-assessment-and-analytical-framework_9789264190511-en.

Patton, M. (2002). *Qualitative research and Publication Methods.* Thousand Oaks,CA: SAGE.

Perkins, D. (2006). Constructivism and troublesome knowledge. In J. Meyer, & R. Land (Eds.), *Overcoming barriers to student uderstanding: threshold concepts and troublesome knowledge* (pp. 33-47). London,UK: Routledge.

Piaget, J. (1929). The child's conception of the world (translated by Joan and Andrew Tomlinson). London, UK: Kegan Paul, Trench, Taubner, & Company.

Poon, C.L. (2014). Five decades of Science Education in Singapore. In A.L. Tan, C.L.Poon, & Lim, S.L. (Eds.), *Inquiry into the Singapore Science Classroom. Research and practices (pp.27-46)*. Singapore: Springer.

Petkova, A. K., & Boyadjieva, P. (1994). The image of the scientist and its functions. *Public Understanding of Science, 3*(2), 215-224.

Random House (2001). *Random House Webster's Unabridged Dictionary (Second Edition)*. USA: Random House.

Roberts, R., & Gott, R. (2002). Investigations: Collecting and using evidence. In D. Sang, & V. Woods-Robinson (Eds.), *Teaching Secondary Scientific Enquiry* (pp. 18-49). London,UK: John Murray.

Roberts, R., & Gott, R. (2003). Assessment of biology investigations. *Journal of Biological Education, 37*(3), 114-121.

Roberts, R. & Gott, R. (2004). A written test for procedural understanding: a way forward for assessment in the UK science curriculum? *Research in Science and Technological Education, 22*(1), 5–21.

Roberts, R. & Gott, R. (2006). Assessment of performance in practical science and pupil attributes. *Assessment in Education, 13*(1), 45–67.

Rollnick, M., Dlamini, B., Lotz, S., & Lubben.F. (2001). Views of South African Chemistry Students in University Bridging Programs on the Reliability of Experimental Data. *Research in Science Education, 31*(4), 553–573.

Ryder, J. (2001). Identifying Science Understanding for Functional Scientific Literacy. *Studies in Science Education, 36*(1), 1-44.

Ryder, J. (2002). Data interpretation activities and students'views of epistemology of science during a University Earth Science Field Study Course. In D. Psillos, & H. Niedderer (Eds.), *Teaching and Llearning in the Science Laboratory*. Dordrecht: Kluwer.

Ryder, J., & Clarke, A. (2001). *Teaching and learning about 'sources of error' on university physics courses.* Retrieved Jan 10, 2010 from http://www- new1.heacademy.ac.uk/ assets/ ps/documents/projects/completed/ undergraduate_understanding_of_the_practices_of_physics.pdf.

Ryder, J., & Leach, J. (1999). University science students experiences of investigative project work and their images of science. *International Journal of Science Education, 21*(9), 945-956.

Ryder, J., & 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.

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Science Education*, *89*(4), 634–656.

Sandoval, W.A., & Reiser, B.J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*(3), 345–372.

Séré, M. (2002). Towards Renewed Research Questions from the Outcomes of the European Project Labwork in Science Education. *Science Education, 86*(5), 624-644.

Séré, M., Fernandez-Gonzalez, M., Gallegos, J.A., Gonzalez-Garcia, F., Manuel, E., Perales, F.J., & Leach, J. (2001). Images of Science Linked to Labwork: A Survey of Secondary School and University Student. *Research in Science Education*, *31*(4), 499–523.

Séré, M., Journeaux, R., & Larcher, C. (1993). Learning the statistical analysis of measurement errors. *International Journal of Science Education, 15*(4), 427 – 438.

Schutt, R.K. (2012*). Investigating the Social World. The Process and Practice of Research*. Thousand Oaks, CA: SAGE.

Shulman, L. (1987). Knowledge and teaching: Foundations of the New Reform. *Harvard Educational Review*, *57*(1), 1–22.

Southerland, S., Smith, M., & Cummins, C. (2000). "What do you mean by that?": Using structured interviews to assess science understanding. In J. Mintzes, J. Wandersee, & J. Novak (Eds.), *Assessing Science Understanding: A human constructivist view* (pp. 72-92). Burlington,MA: Elsevier.

Sharp, J.; Peacock, G., Johnsey, R., Simon, S., Cross, A., & Harris, D. (2012). *Achieving QTS: Meeting the Professional Standards Framework. Primary Science: Teaching Theory and Practice*. Exeter, UK: Learning Matters

Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Thousand Oaks, CA: SAGE

Taylor, J.R. (1997). *Error Analysis. The study of uncertainties in physical measurements*. Sausalito, CA: University Science Books

Tesch, R. (1990). *Qualitative research: Analysis Types and Software Tools*. New York, NY: Falmer.

Toh, K. (1994). Teacher-centred teaching is alive and well. *Teaching and Learning, 15*(1), 12-17.

Tomlinson, J., Dyson, P. J., & Garratt, J. (2001). Student misconceptions of the language of error. *University Chemistry Education,* 5, 1-8.

Varelas, M. (1997). Third and Fourth Graders' Conceptions of Repeated Trials and Best Representatives in Science Experiments. *Journal of Research in Science Teaching, 34*(9), 853-872.

Watson, R. & Wood-Robinson, V. (2002). Investigations: Evaluating evidence. In David Sang and Valerie Wood Robinson (Eds.) *Teaching Secondary Scientific Enquiry*. London, UK: John Murray.

Warwick, P., & Siraj-Blatchford, J. (2006). Using Data Comparison and Interpretation to Develop Procedural Understandings in the Primary Classroom: Case study evidence from action research. *International Journal of Science Education, 28*(5), 443–467.

Wu, H., & Wu, C. (2011). Exploring the Development of Fifth Graders' Practical Epistemologies and Explanation Skills in Inquiry-Based Learning Classrooms. *Research in Science Education, 41*(3), 319-340.

Annex 1.1: Sample of a Type 1 Science Investigation Activity (modified from MOE, 2009)

Slide Along

Additional Information for students:

There is friction when an object moves on surfaces. Friction is a force and it tends to make a moving object slow down and stop.

There is relatively less friction when smooth surfaces rub against each other. The rougher, the greater is the friction.

Friction can be helpful and is often quite necessary. Friction between the soles of our shoes and the ground enables us to walk without slipping. Friction prevents objects from slipping past each other. Without friction, vehicles would not be able to travel on roads because their tires would slip.

However, friction can be disadvantageous as it produces heat and causes things to wear out. Machine parts, soles of our shoes and tires wear out as a result of friction. For this reason, machine parts that rub against one another are either oiled or have ball bearings to reduce friction.

Scenario:

Imagine you have to load boxes of toys onto a truck for them to be delivered to a toy store. To ease your work, find a suitable material for the source of the ramp to facilitate loading and explain why.

Aim:

To find a suitable material for the surface of the ramp that will allow the wooden block to move up with the least force.

Materials:

- A forcemeter
- A wooden block
- A ramp
- Different materials to line the ramp: aluminium foil, sandpaper, plastic sheet, writing paper, and any other possible materials which you want to test out.

Procedure:

1. Draw your experimental set-up
2. Describe your plan

Results:

Types of surfaces on ramp	Amount of force (N)			
	1st reading	2nd reading	3rd reading	Average

Conclusion:

Reflection:

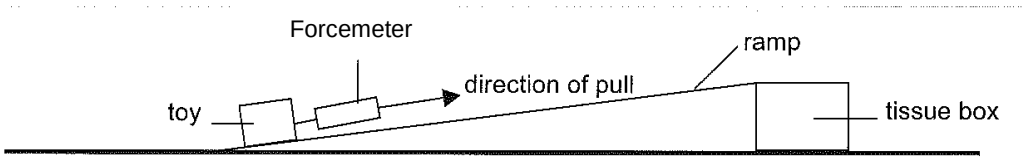
Information for the Teacher:

In this activity, students will set up their investigation to find the surface with the least amount of friction when they move up the toy up the ramp. The amount of friction can be measured using a forcemeter. Students have to record the amount of force that is required to get their toy moving. Students are also encouraged to take the average of at least three readings so as to reduce experimental errors.

The learning outcomes for the investigation are:

(a) To investigate the effect of friction on the motion of objects and communicate findings
(b) To value individual effort and team work

The set-up as shown below is one way students could set up their investigation.



Note: Try out with different possible heights before conducting the investigation with students.

Inform students that it is important that they should pull slowly on the forcemeter with consistent force as they carry out their investigation.

Get them to reflect on what they have learnt on friction:

- How is friction an advantage to us in our daily life?
- How is friction a disadvantage to us in our daily life
- How can I reduce friction?

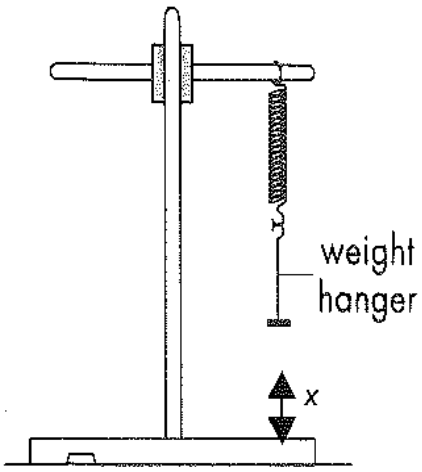
Annex 1.2: Sample of a Type 2 Science Investigation Activity (modified from MOE, 2009)

Spring Along

Scenario:

You are working for a company which has designed a new ride using springs, "Spring Jump". It is important to find out the maximum mass that the spring can hold so that there will be no casualties in this ride.

This can be investigated using a set-up as shown below.



Materials:

- A Retort stand with a spring
- A weight hangar (50g) and weights of different masses
- A ruler

Note that **x** in the set-up should be at least 5 cm above the ground.

Results:

Original Length of spring = _____ cm

Mass of weights (g)	Length of extended spring (cm)	Length of extended spring – original length of spring = Extension of spring (cm)	x (cm)

Conclusion:

Based on the results,

What is the relationship between the mass of weights and the length of the spring?

What is the relationship between the mass of weights and the extension of the spring?

Information for the Teacher:

An object is elastic if it goes back to the original shape and length after being stretched. When an elastic object is stretched, it exerts a force on the object stretching it. An example of an elastic object is the spring. The spring exerts a force called elastic spring force. Elastic spring force is very useful in our daily life. We can find springs in some toys, mattresses and also in forcemeters. However, when a spring is stretched beyond its elastic limit, it will not return to its original length.

The learning outcomes for the investigation are:

(a) To investigate the effects of forces on springs and communicate findings.
(b) To show objectivity by using data and information to validate observations and explanations about forces

Help students to relate the set up to the "Spring Jump". Get students to plan an investigation to find out the maximum mass a spring can take so that when it is extended, it will leave a space of at least 5cm above the ground. The space above the ground (**x**) can be adjusted based on how elastic the springs are.

Get students to first hang a small object on a spring and observe its extension. Lead them to the idea that gravity pulls the object downwards. The object in turn pulls on the spring and causes it to stretch.

Have students measure how much the spring stretches. Introduce the term "extension" and explain that this is the increase in the length of the spring. The value is obtained by subtracting the original length from the length of the stretched spring.

Conclude by getting groups to share their findings with the rest of the class:

- What pattern do you see between the mass of weights and the length/extension of the spring?
- What will happen to the spring if you continue to hang more weights than it can take?

Annex 1.3: Indicators of Skills and Processes in Primary Science (MOE, 2001)

1. **Observing:**
   Pupils should be able to:
   - use all their senses to obtain information about objects and events
   - use instruments to extend the range of the senses and accuracy of the observations
   - notice changes in objects and events
   - identify observations that are relevant to a particular investigation

2. **Comparing**
   Pupils should be able to:
   - identify factors for the purpose of comparison, for example, when comparing a ship and a car, the factors could be function, capacity or cost
   - identify the similarities and differences
   - draw a conclusion about the significance of the similarities or differences

3. **Classifying**
   Pupils should be able to:
   - recognize a common property in a set of objects
   - group a set of objects into two groups based on any one property
   - identify the basis of classification
   - group a set of objects into two or more groups according to one or more common properties
   - identify a common pattern in events or a behavior pattern in organisms
   - generate criteria for grouping
   - use simple classification schemes

4. **Measuring & Using apparatus**
   Pupils should be able to:
   - use measuring devices correctly
   - select appropriate units and instruments
   - exercise care in handling apparatus
   - make estimates and confirm by measuring
   - recognise the variability/reliability of measurement and the need to repeat and check the measurement

5. **Communicating**
   Pupils should be able to:
   - describe an object or an event
   - make a drawing of a given object
   - follow instructions pertaining to an investigation
   - read off information from symbolic representations such as diagrams, tables, bar charts, line graphs, and keys
   - select and present appropriate information in various ways e.g. oral presentation, visual aids, models, electronic documents, multimedia presentation
   - listen to reports/ideas of others and respond to them

6. **Analysing**
   Pupils should be able to:
   - identify parts of a system and the relationship between these parts; and relate the parts to their functions
   - identify patterns and trends in data
   - identify the variable that will affect the investigation
   - identify relationships between the variables
   - identify those aspects of an investigation that make comparisons unfair
   - specify variables to be controlled

7. Generating

Pupils should be able to:

- suggest many, varied and original ideas with some detail
- draw inferences, or conclusions from observations (induction)
- make predictions
- give reasonable explanations based on evidence
- construct hypotheses
- devise ways to test a hypothesis

8. Evaluating

Pupils should be able to:

- decide on the quality and feasibility of an idea or object
- decide whether an inference/hypothesis is supported by observations
- decide on the effectiveness of the method used in an investigation
- construct an idea to explain observations and then test it
- decide on the accuracy of data obtained in an investigation

Annex 2.1 Methods of finding the focal length (Séré et al., 1993)

In the autocollimation method (see Figure 2.26a in Chapter 2), the focal length could be found by first aligning the source of light, the converging lens and a mirror. The light from the source was then passed through the lens and reflected back into the lens by the plane mirror. We would then adjust the distance between the source and the thin lens until a sharp image would be formed on a plane near the source. The distance between the source and thin lens was the focal length. In the Bessel's method, we fixed the distance between the object and the image screen (roughly four times the focal length). We then moved the converging lens between the object and the image screen to find two positions (indicated by $\mathbf{X_L}$ and $\mathbf{X_{L'}}$ in Figure 2.26b in Chapter 2) where sharp images are formed on the screen. The focal length can be determined using the distances **a** and **D** as shown in the figure.

Annex 2.2: Concepts of Evidence (taken from Gott, Duggan, Roberts, & Hussain, 2014)

1 **Fundamental ideas**

0 ***Introduction*** Investigations must be approached with a critical eye. What sort of link is to be established, with what level of measurement and how will opinion and data be weighed as evidence?

This first category pervades the entire scheme and sets the context in which all that follows needs to be judged.

1 ***Opinion and data*** …it is necessary to distinguish between opinion based on scientific evidence and ideas on the one hand, and opinion based on non-scientific ideas (prejudice, whim, hearsay…) on the other.

Distinguishing between the measurable energy emitted from a mobile phone mast and the 'energy' associated with 'crystals'.

2 ***Links*** …a scientific investigation seeks to establish links (and the form of those links) between two or more variables

3 ***Association and causation*** …links can be causal (change in the value of one variable CAUSES a change in another), or associative (changes in one variable and changes in another are linked to some third, and possibly unrecognised, third (or more) variable)

4 ***Types of measurement*** …interval data (measurements of a continuous variable) are more powerful than ordinal data (rank ordering) which are more powerful than categoric data (a label)

Being able to say that 2 wavelengths are 670nm and 460nm is more useful than saying one is longer than the other or that one is red and the other blue.

5 ***Extended tasks*** …measurements, for instance, can be very complicated and constitute a task on their own, but they are only meaningful when set within the wider investigation(s) of which they will form a part

The measurement of 'the absorbancy of a paper towel' involves a 'method' which, all together, contributes to the final measure of absorbancy.

2 **Observation**

0 ***Introduction*** Observation of object and events can lead to informed description and the generation of questions to investigate further.

Observation is one of the key links between the 'real world' and the abstract ideas of science. Observation, in our definition, does not include 'measurement' but rather deals with the way we see objects and events through the prism of our understanding.

1 ***Observing objects*** …objects can be 'seen' differently depending on the conceptual window used to view them.

…a low profile car tyre can be seen as nothing more than that, or it can be seen as a way of increasing the stiffness of the tyre, thus giving more centripetal force with less deformation and thus improving road holding.

2 ***Observing events*** …events can similarly be seen through different conceptual windows.

…the motion of a parachute is seen differently when looked at through a framework of equal and unequal forces and their corresponding accelerations.

3 ***Using a key*** … the way in which an object can be 'seen' can be shaped by using a key

e.g. a branching key gives detailed clues as to what to 'see'. It is, then, a heavily guided concept-driven observation

4 ***Taxonomies*** …taxonomies are a means of using conceptually driven observations to set up classes of objects or organisms that exhibit similar/different characteristics or properties with a view to using the classification to solve a problem.

....organisms observed in a habitat may be classified according to their feeding characteristics (to track population changes over time for instance) or a selection of materials classified into efficient conductors identified from inefficient conductors

5 ***Observation and experiment*** …observation can be the start of an investigation, experiment of survey.

…noticing that shrimp populations vary in a stream leads to a search for a hypothesis as to why that is the case, and an investigation to test that hypothesis.

6 ***Observation and map drawing*** … technique used in biological and geological fieldwork to map a site based on conceptually driven observations that illustrate features of scientific interest

....an ecologist may construct a map of a section of a stream illustrating areas of varying stream flow rate or composition of the stream bed.

3 **Measurement**

0 ***Introduction*** Measurement must take into account inherent variation due to uncontrolled variables, human error and the characteristics of the instruments used.

This section lies at the very centre of our model for measurement, data and evidence and is fundamental to it.

1 ***Inherent variation*** …the measured value of any variable will never repeat unless all possible variables are controlled between measurements - circumstances which are very difficult to create

Repeated bounces of a squash ball under ostensibly identical conditions will result in varied data.

2 ***Human error*** …the measured value of any variable can be subject to human error which can be random, or systematic

In the case of the squash ball, human error could result from a shaky hand when the ball was released (random) or the bounce height could always appear higher than it really is if the observer was below the height looking up at the bounce against the rule

4 Instruments: underlying relationships

1 ***Linear relationships*** ...most instruments rely on an underlying and preferably linear relationship between two variables.

e.g. A thermometer relies on the relationship between the volume of a liquid and temperature.

2 ***Non-linear relationships*** ...some 'instruments', of necessity, rely on non-linear relationships.

e.g. Moving iron ammeter, pH

3 ***Complex relationships*** ...the relationship may not be straightforward and may be confounded by other factors.

e.g. The prevalence, or size, of a species of lichen is an indicator of the level of pollution but other environmental factors such as aspect, substrate, or air movement can also affect the distribution of lichen.

4 ***Multiple relationships*** ...sometimes several relationships are linked together so that the measurement of a variable is indirect.

e.g. Medical diagnosis often relies on indirect multiple relationships. Also, braking distance is an indirect measure of frictional force.

5 Instruments: calibration and error

0 ***Introduction*** Instruments must be carefully calibrated to minimise the inevitable uncertainties in the readings

All instruments must be calibrated so that the underlying relationship is accurately mapped onto the scale. If the relationship is non-linear, the scale has to be calibrated more often to map that non-linearity. All instruments, no matter how well made, are subject to error. Each instrument has finite limits on, for example, its resolution and sensitivity.

1 ***End points*** ...the instrument must be calibrated at the end points of the scale.

e.g. A thermometer must be calibrated at 0 °C and 100 °C.

2 ***Intervening points*** ...the instrument must be calibrated at points in between to check the linearity of the underlying relationship.

e.g. A thermometer must be calibrated at a number of intervening points to check, for instance, for non-linearity due to non-uniform bore of the capillary.

3 ***Zero Errors*** ...there can be a systematic shift in scale and that instruments should be checked regularly.

e.g. If the zero has been wrongly calibrated, if the instrument itself was not zeroed before use, or if there is fatigue in the mechanical components, a systematic error can occur.

4 ***Overload, limiting sensitivity / limit of detection*** ...there is a maximum (full scale deflection) and a minimum quantity which can be measured reliably with a given instrument and technique.

e.g. change in mass when Mg burns in air could not be detected on scales that measure to only whole grams.

5 ***Sensitivity**** ...the sensitivity of an instrument is a measure of the amount of error inherent in the instrument itself.

e.g. An electronic voltmeter will give a reading that fluctuates slightly.

6 ***Resolution and error*** ...the resolution is the smallest division which can be read easily. The resolution can be expressed as a percentage.

e.g. If the instrument can measure to 1 division and the reading is 10 divisions, the error can be expressed as 10±1 or as a percentage error of 10%.

7 ***Specificity*** ...an instrument must measure only what it purports to measure.

e.g. false positives on drug tests due to detection of a similar naturally occurring substance.

8 ***Instrument use*** ...there is a prescribed procedure for using an instrument which, if not followed, will lead to systematic and/or random errors.

e.g. When measuring the temperature of a liquid, if one takes the thermometer out of the liquid to read the thermometer, this will lead to systematically low or high readings, compared to reading the thermometer immersed in the liquid. More specifically, there is a prescribed depth of immersion for some thermometers that takes account of the expansion or contraction of the glass and the mercury (or alcohol) that are not in the liquid being measured.

9 ***Human error*** ...even when an instrument is chosen and used appropriately, human error can occur.

e.g. Scales on measuring instruments can easily be misread.

6 **Reliability and validity of a single measurement**

0 ***Introduction*** Any measurement must be reliable and valid.

A measurement, once made, must be scrutinised to make sure that it is a valid measurement; it is measuring what was intended, and that it can be relied upon. Repeating readings and triangulation, by using more than one of the same type of instrument or by using another type of instrument, can increase reliability.

1 ***Reliability*** …a reliable measurement requires an average of a number of repeated readings; the number needed depends on the accuracy required in the particular circumstances

e.g. the height from which a ball is dropped could be checked if it was important that the drop height was accurate.

2 ***Reliability*** ...instruments can be subject to inherent inaccuracy so that using different instruments can increase reliability.

e.g. Measurement of blood alcohol level can be assessed with a breathalyser and cross checked with a blood test. Also, temperature can be measured with mercury, alcohol, and digital thermometers to ensure reliability.

3 ***Reliability*** ...human error in the use of an instrument can be overcome by independent, random checks.

e.g. Spot checks of measurement techniques by co-workers are sometimes built into routine procedures.

4 ***Validity*** ...measures that rely on complex or multiple relationships must ensure that they are measuring what they purport to measure.

e.g. is the colour change a measure of bacterial activity or might something else have caused it?

7 **The choice of an instrument for measuring a datum**

0 ***Introduction*** Measurements are never entirely accurate for a variety of reasons.

Of prime importance is choosing the instrument to give the accuracy and precision required; a proactive choice rather than a reactive discovery that it wasn't the right instrument for the job!

1 ***Trueness or accuracy**** ...trueness is a measure of the extent to which repeated readings of the same quantity give a mean that is the same as the 'true' mean.

e.g. If the mean of a series of readings of the height of an individual pupil is 173 cm and her 'true' height, as measured by a clinic's instrument is 173 cm, the measuring instrument is 'true' .

2 ***Non-repeatability*** ...repeated readings of the same quantity with the same instrument never give exactly the same answer.

e.g. Weighing yourself on a bathroom scale in different places on the bathroom floor, or standing in a slightly different position on the scales, will result in slightly differing measurements. It is never possible to repeat the measurement in exactly the same way.

3 ***Precision*** … (Sometimes called "imprecision" in industry) refers to the observed variations in repeated measurements from the same instrument. In other words, precision is an indication of the spread of the repeated measurements around the mean. A precise measurement is one in which the readings cluster closely together. The less the instrument's precision, the greater is its uncertainty. A precise measurement may not necessarily be an accurate or true measurement (and vice versa). The concept of precision is also called "reliability" in some fields. A more formal descriptor or assessment of precision might be the range of the observed readings, the standard deviation of those readings, or the standard error of the instrument itself.

e.g. For bathroom scales, a precise set of measurements might be: 175, 176, 175, 176, and 174 pounds.

4 ***Reproducibility*** …whereas repeatability (precision) relates to the ability of the method to give the same result for repeated tests of the same sample on the same equipment (in the same laboratory), reproducibility relates to the ability of the method to give the same result for repeated tests of the same sample on equipment in different laboratories.

e.g. 'Round Robins' are often used to check between different laboratories. A standardised sample is sent to each lab and they report their measurement(s) and degree of uncertainty. Labs are then compared.

5 ***Outliers in relationships*** …outliers, aberrant or anomalous values in data sets should be examined to discover possible causes. If an aberrant measurement or datum can be explained by poor measurement procedures (whatever the source of error), then it can be deleted.

e.g. an anomalous bounce would be deleted if the cause of the anomaly was known, but if it could not be explained, and then further bounces would be needed to see if it was part of the inherent variation.

8 **Sampling a datum**

0 ***Introduction*** A series of measurement of the same datum can be used to determine the reliability of the measurement

We use the term 'sampling' to mean any sub-set of a population. The population might be a species of animal or plant, or even the possible sites where gold might be found. We shall also take the population to mean the infinite number of repeated readings.

1 ***Sampling*** …one or more measurements comprise a sample of all the possible measurements that could be made.

e.g. The measurement of a single blade of grass is a sample of all the blades of grass in a field. Also, a single measurement of the bounce height of a ball is a sample of the infinite number of such bounces that could be measured.

2 ***Size of sample*** …the number of measurements taken. The greater the number of readings taken, the more likely they are to be representative of the population.

e.g. repeated readings on a ammeter in a particular circuit are a sample of all possible readings. The more readings taken, the more the sample represents the population of all possible readings.

3 ***Reducing bias in sample/representative sampling*** …measurements must be taken using an appropriate sampling strategy, such as random sampling, stratified or systematic sampling so that the sample is as representative as possible.

To find the height of college students, tables of random numbers can be used to select students.

4 ***An anomalous datum*** …an unexpected datum could be indicative of inherent variation in the data or the consequence of a recognised uncontrolled variable

e.g. Continuing the above examples, a very small height may have been recorded from a child visiting the college and should not be part of the population being sampled; whereas a very low rebound height from a squash ball may occur as a result of differences in the material of the ball and is therefore part of the sample.

9 **Statistical treatment of measurements of a single datum**

0 ***Introduction*** A group of measurements of the same datum can be described in various mathematical ways.

The statistical treatment of a datum is concerned with the probability that a measurement is within certain limits of the true measurement. The following are some basic statistics associated with a single datum.

1	***Range***	…the range is a simple description of the distribution and defines the maximum and minimum values measured. *e.g. Measuring the height of carbon dioxide bubbles on successive trials in a yeast experiment, the following measurements were recorded and ordered sequentially: 2.7, 2.9, 3.1, 3.1, 3.1, 3.3, 3.4, 3.4, 3.5, 3.6 and 3.7 cm. The range is 1.0 cm (3.7 - 2.7).*
2	***Mode***	...the mode is the value which occurs most often. *e.g. Continuing the example above, the mode is 3.1 cm.*
3	***Median***	...the median is the value below and above which there are half the measurements. *e.g. Continuing the example above, the median is 3.3 cm.*
4	***Mean***	...the mean (average) is the sum of all the measurements divided by the number of measurements. *e.g. Continuing the example above, the mean is 3.2 cm*
5	***Frequency distributions***	...a series of readings of the same datum can be represented as a frequency distribution by grouping repeated measurements which fall within a given range and plotting the frequencies of the grouped measurements.
6	***Standard deviation***	…the standard deviation (SD) is a way of describing the spread of normally distributed data. The standard deviation indicates how closely the measurements cluster around their mean. In other words, the standard deviation is a measure of the extent to which measurements deviate from their mean. The more closely the measurements cluster around the mean, the smaller the standard deviation. The standard deviation depends on the measuring instrument and technique - the more precise these are, the smaller the standard deviation of the sample or of repeated measurements. *e.g. Continuing the example above, SD = 0.30 cm.*
7	***Standard deviation of the mean (standard error)***	…the standard deviation of the mean describes the frequency distribution of the means from a series of readings repeated many times. The standard deviation of the mean depends on the measuring instrument and technique and on the number of repeats. The standard error of measurement is an estimate of the probable range within which the' true' mean falls; that is, an estimate of the uncertainty associated with the datum *e.g. Continuing the example above, SE = 0.09 cm*.
8	***Coefficient of variation***	...the coefficient of variation is the standard deviation expressed as a percentage of the mean (CV = SD*100/mean). *e.g. Continuing the example above, the coefficient of variation is 9.4%*

9 ***Confidence limits*** ...confidence limits indicate the degree of confidence that can be placed on the datum. For example, '95% confidence limits' means that the 'true' datum lies within 2 standard errors of the calculated mean, 95% of the time. Similarly '68% confidence limits' means that the 'true' datum lies within 2 standard errors of the calculated mean, 68% of the time.

e.g. Continuing the example above, the true value of the datum lies within 0.18 cm (2 standard errors) of 3.2 cm (the mean), 19 times out of 20. The upper and lower confidence limits at the 95% level are 3.38 (3.2 + 0.18) and 3.02 (3.2 - 0.18) respectively. In other words, the 'true' value lies between 3.02 and 3.38 cm, 95% of the time.

10 **Reliability and validity of a datum**

0 ***Introduction*** A datum must have a known (or estimated) reliability and validity before it can be used in evidence.

Any datum must be subject to careful scrutiny to ascertain the extent to which it:

- is valid: that is, has the value of the appropriate variable been measured? Has the parameter been sampled so that the datum represents the population?

- is reliable: for example, does the datum have sufficient precision?

The wider the confidence limits (the greater the uncertainty), the less reliable the datum. Only then can the datum be weighed as evidence. Evaluation a datum also includes evaluating the validity of the ideas associated with the making of a single measurement.

1 ***Reliability*** ...a datum can only be weighed as evidence once the uncertainty associated with the instrument and the measurement procedures have been ascertained.

The reliability of a measurement of blood alcohol level should be assessed in term s of the uncertainty associated with the breathalyser (e.g. +/- 0.01) and in terms of how the measurement was taken (e.g. superficial breathing versus deep breathing).

2 ***Validity*** ...that a measurement must be of, or allow a calculation of, the appropriate datum.

The girth of a tree is not a valid indicator of the tree's age.

11 **Design of investigations: Variable structure**

0 ***Introduction*** The design of an investigation requires variables to be identified (as Independent, dependent and controlled) and measured.

An investigation is an attempt to determine the relationship, or lack of one, between the independent and dependent variables, or between two or more sets of data. Investigations take many forms but all have the same underlying structure. By identifying and understanding the basic structure of an investigation in terms of variables and types of variables, we can begin to evaluate the

1 ***The independent variable*** ...the independent variable is the variable for which values are changed or selected by the investigator.

e.g. The 'type of ball' used in an investigation to compare the bounciness of different types of balls. Also, the 'depth in a pond' at which light intensity is to be measured.

2 ***The dependent variable*** ...the dependent variable is the variable the value of which is measured for each and every change in the independent variable.

e.g. Continuing the examples above, the height to which each type of ball bounces. Also, the light intensity at each of the chosen depths in the pond.

3 ***Correlated variables*** …in some circumstances we are looking for a correlation only rather than any implied causation

e.g. Foot size can be predicted from hand size (both caused by other factors).

4 ***Categoric variables*** ...a categoric variable has values which are described by labels. Categoric variables are also known as nominal data.

e.g. The variable "type of metal" has data values of "iron", "copper", etc.

5 ***Ordered variables*** ...an ordered variable has values which are also descriptions, labels or categories but these categories can be ordered or ranked. Measurement of ordered variables results in ordinal data.

e.g. The variable of size e.g.' very small', 'small', 'medium ' or 'large' is an ordered variable. Although the labels can be assigned numbers (e.g. very small=1, small=2 etc.) size remains an ordered variable.

6 ***Continuous variables*** ...a continuous variable is one which can have any numerical value and its measurement results in interval data.

e.g. Weight, length, force.

7 ***Discrete variables*** ...a discrete variable is a special case in which the values of the variable are restricted to integer multiples.

e.g. The number of discrete layers of roof insulation.

8 ***Multivariate designs*** ...a multivariate investigation is one in which there is more than one independent variable.

e.g. The effect of the width and the length of a model bridge on its strength. Also, the effect of temperature and humidity on the distribution of gazelles in a particular habitat.

12 Design: Validity, 'fair tests' and controls

0 ***Introduction*** Uncontrolled variation can be reduced through a variety of techniques.

Fair tests and controls aim to isolate the effect of the independent variable on the dependent variable. Laboratory-based investigations, at one end of the spectrum, involve the investigator changing the independent variable and keeping all the controlled variables constant. This is often called 'the fair test', but it is no more than one of several valid designs. At the other end of the spectrum are field studies where many naturally changing variables are measured and correlations sought. For example, an ecologist might measure many variables in a habitat over a period of time. Having collected the data, correlations might be sought between variables such as day length and emergence of a butterfly, using statistical treatments to ensure validity. The possible effect of other variables can be reduced by only considering data where the values of other variables are the same or similar. In between these extremes, there are many types of valid designs that involve different degrees of manipulation and control. Fundamentally, all these investigations have a similar structure; what differ are the strategies to ensure validity.

1 ***Fair test*** ...a fair test is one in which only the independent variable has been allowed to affect the dependent variable.

e.g. A laboratory experiment about the effect of temperature on dissolving time, where only the temperature is changed. Everything else is kept exactly the same.

2 ***Control variables in the laboratory*** ...other variables can affect the results of an investigation unless their effects are controlled by keeping them constant.

e.g. In the above experiment, the mass of the chemical, the volume of liquid, the stirring technique, and the room temperature are some of the variables that should be controlled.

3 ***Control variables in field studies*** ...some variables cannot be kept constant and all that can be done is to make sure that they change in the same way.

e.g. In a field study on the effect of different fertilisers on germination, the weather conditions are not held constant but each experimental plot is subjected to the same weather conditions. The conditions are matched.

4 ***Control variables in surveys*** …the potential effect on validity of uncontrolled variables can be reduced by selecting data from conditions that are similar with respect to other variables.

e.g. In a field study to determine whether light intensity affects the colour of dog's mercury leaves, other variables are recorded, such as soil nutrients, pH and water content. Correlations are then sought by selecting plants growing where the value of these variables is similar.

5 ***Control group experiments*** …control groups are used to ensure that any effects observed are due to the independent variable(s) and not some other unidentified variable. They are no more than the default value of the independent variable.

e.g. In a drug experiment, patients with the same illness are divided into an experimental group who are given the drug and a control group who are given a placebo or no drug.

13 Design: Choosing values

0 ***Introduction*** Choosing the values of the variables in an investigation.

The values of the variables need to be chosen carefully. This is possible in the majority of investigations, prior to the data being collected. In field studies where data are collected from variables that change naturally, some of these concepts can only be applied retrospectively.

1 ***Trial run*** …a trial run can be used to establish the broad parameters required of the experiment (scale, range, number) and help in choosing instrumentation and other equipment

e.g. Before drug experiments are carried out, trials are conducted to determine appropriate dosage and appropriate measures of side effects, among other things.

2 ***The sample*** ...issues of sample size and representativeness apply in the same way as in sampling a datum (see Measuring a datum, 2).

e.g. The choice of sample size and the sampling strategy will directly affect the validity of the findings.

3 ***Relative scale*** ...the choice of sensible values for quantities is necessary if measurements of the dependent variable are to be meaningful.

e.g. In differentiating the dissolving times of different chemicals, a large quantity of chemical in a small quantity of water causing saturation will invalidate the results.

4 ***Range*** ...the range over which the values of the independent variable is chosen is important in ensuring that any pattern is detected.

e.g. An investigation into the effect of temperature on the volume of yeast dough using a range of 20 to 25 °C would show little change in volume.

5 ***Interval*** ...the choice of interval between values determines whether or not the pattern in the data can be identified.

e.g. An investigation into the effect of temperature on enzyme activity would not show the complete pattern if 20°C intervals were

6 ***Number*** ...a sufficient number of readings is necessary to determine the pattern.

e.g. The number is determined partly by the range and interval issues discussed above, but in some cases for the complete pattern to be seen, more readings may be necessary in one part of the range than another. This applies particularly if the pattern changes near extreme values, for example, in a spring extension experiment at the top of the range of the mass suspended on the spring.

14 **Design: Accuracy and precision**

0 ***Introduction*** Ensuring appropriate accuracy and precision.

The design of the investigation must provide data with sufficiently appropriate accuracy and precision to answer the research question. This consideration should be built into the design of the investigation. Different investigations will require different levels of accuracy and precision depending on their purpose.

1 ***Determining differences*** ...there is a level of precision which is sufficient to provide data which will allow discrimination between two or more means.

e.g. The degree of precision required to discriminate between the bounciness of a squash ball and a ping pong ball is far less than that required to discriminate between two ping pong balls.

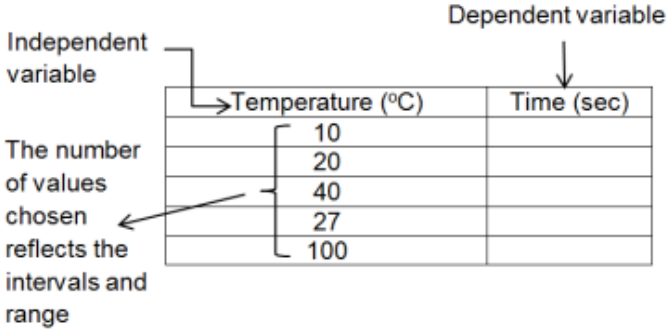
2 ***Determining patterns*** ...there is a level of precision which is required for the trend in a pattern to be determined.

e.g. Large error-of-measurement bars on the points of a line graph may not allow discrimination between an upward curve or a straight line.

15 **Design: Tables**

1 ***Tables*** ...tables can be used as organisers for the design of an experiment by preparing the table in advance of the whole experiment. A table has a conventional format.

e.g. An experiment on the effect of temperature on the dissolving time of calcium chloride:



16 **Reliability and validity of the design**

0 ***Introduction*** An evaluation of an investigation must consider reliability and validity.

In evaluating the design of an investigation, there are two overarching questions:

- will the measurements result in sufficiently reliable data to answer the question?

- will the design result in sufficiently valid data to answer the question?

Evaluation the design of an investigation included evaluating the reliability and validity of the ideas associated with the making of a single measurement and with each and every datum.

1 ***Reliability of the design*** ...the reliability of the design includes a consideration of all the ideas associated with the measurement of each and every datum.

e.g. Factors associated with the choice of the measuring instruments to be used must be considered, for instance, the error associated with each measuring instrument. The sampling of each datum and the accuracy and precision of the measurements should also be considered. This includes the sample size, the sampling technique, relative scale, the range and interval of the measurements, the number of readings, and the appropriate accuracy and precision of the measurements.

2 ***Validity of the design*** ...the validity of the design includes a consideration of the reliability (as above) and the validity of each and every datum.

e.g. This includes the choice of measuring instrument in relation to whether the instrument is actually measuring what it is supposed to measure. This includes considering the ideas associated with the variable structure and the concepts associated with the fair test. For instance, measuring the distance travelled by a car at different angles of a ramp will not answer a question about speed as a function of angle.

17 **Data presentation**

0 ***Introduction*** Data can be presented in a number of ways.

Having established that the design of an investigation is reliable and valid, what do we need to understand to explore the relationship between one variable and another? Another way of thinking about this is to think of the pattern between two variables or two sets of data. What do we need to understand to know that the pattern is valid and reliable? The way that data are presented allows patterns to be seen. There is a close link between graphical representations and the type of variable they represent.

1	***Tables***	…a table is a means of reporting and displaying data. But a table alone presents limited information about the design of an investigation e.g. control variables or measurement techniques are not always overtly described. *e.g. Simple patterns such as directly proportional or inversely proportional relationships can be shown effectively in a table.*
2	***Bar charts***	...bar charts can be used to display data in which the independent variable is categoric and the dependent variable is continuous. *e.g. The number of pupils who can and cannot roll their tongues would be best presented on a bar chart.*
3	***Line graphs***	...line graphs can be used to display data in which both the independent variable and the dependent variable are continuous. They allow interpolation and extrapolation. *e.g. The length of a spring versus the force applied would be best displayed in a line graph.*
4	***Scatter graphs (or scatter plots)***	...are used to display data in which both the independent variable and the dependent variable are continuous. Scatter graphs are often used where there is much fluctuation in the data because they can allow an association to be detected. Widely scattered points can show a weak correlation, points clustered around, for example, a line can indicate a relationship. *e.g. The dry mass of the aerial parts of a plant and the dry mass of the roots.*
5	***Histograms***	...histograms can be used to display data in which a continuous independent variable has been grouped into ranges and in which the dependent variable is continuous. *e.g. On a seashore, the distance from the sea could be grouped into ranges and the number of limpets in each range plotted in a histogram*
6	***Box and whisker plots***	...the box, in box and whisker plots, represents 50% of the data limited by the 25th and 75th percentile. The central line is the median. The limits of the 'whiskers' may show either the extremes of the range or the 2.5% and 97.5% values. *e.g. Box and whisker plots are often used to compare large data sets.*
7	***Multivariate data***	…nested or multiple tables, 3D bar charts and line graphs (surfaces) are suitable for some forms of multivariate data.
8	***Other forms of display***	...data can be transformed, for example, to logarithmic scales so that they meet the criteria for normality which allows the use of parametric statistics. *e.g. Logarithmic transformation is commonly used in clinical and laboratory medicine, weather maps etc.*

18 Statistics for analysis of data

0 ***Introduction*** Statistical techniques can be used to analyse data.

Statistical techniques used for analysing data address three main questions:

- do the two groups of data differ from each other (by probabilistic chance alone?)?

- do data change when repeated measurements are taken on a second separate occasion?

- is there an association between row sets of data?

Statistics consider the variability of the data and presents a result based on probability. Each statistical technique has associated criteria depending on, for example, the type of data, its distribution, sample size, etc. Some common methods used in the statistical analysis of data are described here.

1 ***Differences between means*** …a t-test can be used to estimate the probability that two means from normally distributed populations, derived from an investigation involving a categoric independent variable, are different (i.e. what is the chance that the two means probably occurred by chance alone?). If measures are repeated with the same or matched pairs, then a paired t-test can be used.

2 ***Analysis of variance*** ...analysis of variance is a technique which can be used to estimate the effects of a number of variables in a multivariate problem involving categoric independent variables.

3 ***Linear and non-linear regression*** ...regression can be used to derive the 'line of best fit' for data resulting from an investigation involving a continuous independent variable.

4 ***Non-parametric measures*** ...when the measurements are not normally distributed, nonparametric tests, such as the Mann-Whitney U-test, can be used to estimate the probability of any differences.

5 ***Categoric data*** ...when the data results from an investigation in which both independent and dependent variables are categoric, the analysis of the data must use, for instance, a chi-squared test.

19 Patterns and relationships in data

0 ***Introduction*** Data must be inspected for underlying patterns.

Patterns cannot be treated in isolation from the physical system that they represent, because patterns represent the behaviour of variables in that system. Patterns can be seen in tables or graphs or can be reported by using the results of appropriate statistical analysis. The interpretation of patterns and relationships must respect the limitations of the data. For instance, there is a danger of over-generalizing or of implying causality when there may be a different, less direct type of association.

1 ***Types of patterns*** …there are different types of association such as causal, consequential, indirect or chance associations. “Chance association” means that observed differences in data sets, or changes in data over time, happen simply by chance alone. We must sceptically be open to possibility that a pattern has emerged by chance alone. Statistical tests give us a rational way to estimate this chance.

e.g. In any large multivariate set of data, there will be associations, some of which will be chance associations. Even if x and y are highly correlated, x does not necessarily cause y: y may cause x or z may cause x and y.

(See 1.3.) Also, changes in students' understanding before and after an intervention may not be significant and/or may be due to other factors.

2 ***Linear relationships*** ...straight line relationships (positive slopes, negative, and vertical and horizontal as special cases) can be present in data in

tables and line graphs and that such relationships have

important predictive power (y = mx + c)

e.g. Height and time for a falling object.

3 ***Proportional relationships*** …direct proportionality is a particular case of a straight line relationships with consequent predictive characteristics. The relationship is often expressed in the form (y = mx).

e.g. Hooke's law: the length of a spring is directly proportional to the force on the spring.

4 ***'Predictable' curves*** …patterns can follow predictable curves (y=x2 for instance), and that such patterns are likely to represent significant regularities in the behaviour of the system (velocity against time for a falling object for instance)

e.g. Velocity against time for a falling object. Also, the terminal velocity of a parachute against its surface area.

5 ***Complex curves*** .. some patterns can be modelled mathematically to give approximations to different parts of the curve (Hooke's law for a spring taken beyond its elastic limit for instance)

e.g. Hooke's law for a spring taken beyond its elastic limit.

6 ***Empirical relationships*** ... patterns can be purely empirical and not be easily represented by any simple mathematical relationship (traffic flow as a function of time of day for instance)

e.g. Traffic flow as a function of time of day.

7 ***Anomalous data*** … patterns in tables or graphs can show up anomalous data points which require further consideration before excluding them from further consideration (the 'bad' measurement due to human error perhaps)

e.g. A 'bad' measurement or datum due to human error.

8 ***Line of best fit*** …for line graphs (and scatter graphs in some cases) a 'line of best fit' can be used to illustrate the underlying relationship, 'smoothing out' some of the inherent (uncontrolled) variation and human error

20 **Reliability and validity of the data in the whole investigation**

0 ***Introduction*** An overall solution to a problem can included repeated experiments and triangulation from other data sources.

So far we have considered the data within a single investigation. In reality the results of an investigation will usually be compared with evidence from other investigations.

In evaluating the whole investigation, all the foregoing ideas about evidence need to be considered in relation to the two overarching questions:

- are the data reliable?

- are the data valid?

In addressing these two questions, ideas associated with the making of single measurements and with each and every datum in an investigation should be considered. The evaluation should also include a consideration of the design of an investigation, as well as ideas associated with measurement, with the presentation of data, and with the interpretation of patterns and relationships.

1 ***A series of experiments*** ...a series of experiments can add to the reliability and validity of evidence even if, individually, their precision does not allow much weight to be placed on the results of any one experiment alone.

2 ***Secondary Data*** ...data collected by others is a valuable source of additional evidence, provided its value as evidence can be judged.

e.g. Meta-analysis

3 ***Triangulation*** ...triangulation with other methods can strengthen the validity of the evidence.

21 **Relevant societal issues**

0 ***Introduction*** Evidence must be considered in the light of personal and social experience and the status of the investigators.

If we are faced with evidence and we want to arrive at a judgement or decision that leads to action, other factors outside the domain of science may become relevant, some of which are listed here.

1 ***Credibility of evidence*** ...credibility has a lot to do with face validity: consistency of the evidence with conventional ideas, with common sense, and with personal experience. Credibility increases with the degree of scientific consensus on the evidence or on theories that support the evidence. Credibility can also turn on the type of evidence presented, for instance, statistical versus anecdotal evidence.

e.g. Evidence showing low emissions of dioxins from a smokestack is compromised by photos of black smoke spewing from the smokestack (even though dioxins are relatively colourless). Also, concern for potential health hazards for workers in some industries often begins with anecdotal evidence, but is initially rejected as not being scientifically credible.

2 ***Practicality of consequences*** ...the implications of the evidence may be practical and cost effective, or they may not be. The more impractical or costly the implications, the greater the demand for higher standards of validity and reliability of the evidence.

e.g. The negative side effects of a drug may outweigh its benefits, for all but terminally ill patients. Also, when judging the evidence on the source of acid rain, Americans will likely demand a greater degree of certainty of the evidence than Canadians who live down wind, because of the cost to American industries to reduce sulphur

3 ***Experimenter bias*** ...evidence must be scrutinized for inherent bias of the experimenters. Possible bias may be due to funding sources, intellectual rigidity, or an allegiance to an ideology such as scientism, religious fundamentalism, socialism, or capitalism, to name but a few. Bias is also directly related to interest: Who benefits? Who is burdened?

e.g. Studying the link between cancer and smoking funded by the tobacco industry; or studying the health effects of genetically modified foods funded by Greenpeace. Also, the acid rain issue (above) illustrates different interests on each side of the Canadian/American border.

4 ***Power structures*** ...evidence can be accorded undue weight, or dismissed too lightly, simply by virtue of its political significance or due to influential bodies. Trust can often be a factor here. Sometimes people are influenced by past occurrences of broken trust by government agencies, by industry spokespersons, or by special interest groups.

e.g. Studies published in the New England Journal of Medicine tend to receive greater weight than other studies. Also, the pharmaceutical industry's negative reaction to Dr. Olivieri's research results that were not supportive of their drug Apotex at Toronto's Hospital for Sick Children in 2001

5 ***Paradigms of practice*** ...different investigators may work within different paradigms of research. For instance, engineers operate from a different perspective than scientists. Thus, evidence garnered within one paradigm may take on quite a different status when viewed from another paradigm of practice.

e.g. Theoretical scientists tend to use evidence to support arguments for advancing a theory or model, whereas scientists working for an NGO, for instance, tend to use evidence to solve a problem at hand within a short time period. Theoretical scientists have the luxury of subscribing to higher standards of validity and reliability for their evidence.

6 ***Acceptability of consequences*** …Evidence can be denied or dismissed for what may appear to be illogical reasons such as public and political fear of its consequences. Prejudice and preconceptions play a part here.

e.g. During the tainted blood controversies in the mid-1980s, the Canadian Red Cross had difficulty accepting evidence concerning the transmission of HIV in blood transfusions. BSE and traffic pollution are examples in Europe.

7 ***Status of experimenters*** ...the academic or professional status, experience and authority of the experimenters may influence the weight which is placed on the evidence.

e.g. Nobel laureates may have their evidence accepted more easily than new researchers' evidence. Also, A botanist's established reputation affects the credibility of his or her testimony concerning legal evidence in a courtroom.

8 ***Validity of conclusions*** …conclusions must be limited to the data available and not go beyond them through inappropriate generalisation, interpolation or extrapolation

e.g. The beneficial effects of a pharmaceutical may be limited to the population sample used in the human trials of the new drug. Also, evidence acquired from a male population concerning a particular cardiac problem may not apply as widely to a female population.

Annex 3.1: Interview 1

Aim: to understand different conceptions about variation in repeated readings i.e. causes of variation in repeated readings; instruments and the quality of the data; students' tendency to put variation all down to 'human error'; distinguishing patterns in data with variation.

Probe 1:

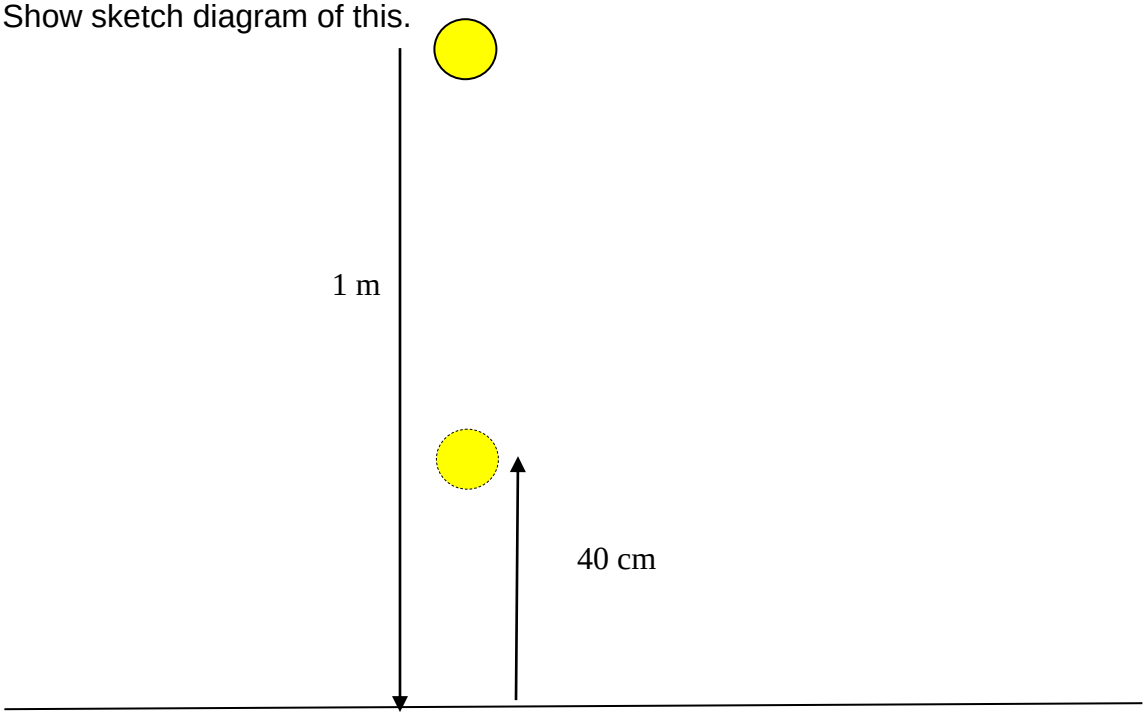
Q1. Scientists usually repeat readings if possible, rather than just taking one. Why do they repeat readings?

- If response is 'because that's what scientists do', ask again why scientist might do this.
- If responses is about providing practice to improve the process of taking readings, ask again whether this means the scientist is looking for a perfect reading (a 'true' value)
- If responses are about 'checking': ask what they might be 'checking' for.
- If responses are 'to get a mean'; ask why a mean might be necessary. And, what's so special about a 'mean'/what is the importance of a mean?
- If response is about to improve accuracy, ask what do they mean by accurate readings, is it about getting close to a 'true' value?
- If response is about getting a spread, ask why would scientist think that a spread is essential?

Probe 2:

Imagine a squash ball was dropped from the height of one metre and you measured its rebound height against the metre rule. It bounced back to 40.0 cm.

Show sketch diagram of this.



Give student the table below:

Student name: ALREADY FILLED IN

Bounce height in cm										
40.0										

Q2. If you were to drop the ball again, what reading do you think you'd get? Fill in the next space in the table. [Prompt: Same? Different?] Why? What made you write that value?

Q3. Imagine you bounced it 10 times. What would the other 8 readings look like? Fill in imaginary results into the table. Why do you think they'd be like this? What does your data show?

Q4. What about 50, or 100 times? Explain.

- If responses show reducing or no variation with more repeats, ask why the values seem to get more and more alike as the number of repeats increases? Would they be surprised if a [suggest a 'more varied' value] appeared again? Why?

- If repeated readings end up with no variation, ask ‘why’? Or ‘can you explain?’
- If no reduction in variation/random values, ask ‘why’? Or ‘can you explain?’ Wouldn’t you expect the results to get more similar?
- If all the explanations are about ‘human error’, ask if there might be any other reason their results vary or ‘get better’ (if they’ve used words like this in their explanation).

Q5. If instead of measuring the rebound height by eye against the ruler you’d used a video and had ‘freeze framed’ it at the highest point, would you have got the same results (as the 10 in the table)? Why? If they were different, how would they differ? Why do you think that?

Q6. Do you think there is such a thing as a ‘perfect’ measuring instrument? What would a ‘perfect’ measuring instrument be like? Is it ‘error-free’? Or, would such an instrument be expected to give you the same reading over and over again?

Probe 3:

A class of 12 pupils did an osmosis experiment. All did the same experiment, using both apple and potato, and pooled/collated their results on the board. They placed potato and apple 'chips' of equal size and mass in a sugar solution and measured the change in mass of the 'chips' on a top pan balance and recorded the results as a percentage of the original mass after 4 hours.

Their results are in the table below:

Percentage of original mass												
Apple	105.03	104.49	107.38	104.69	107.36	105.63	105.25	104.37	102.97	99.02	104.69	104.77
Potato	95.24	95.31	94.00	99.43	95.17	93.42	94.73	94.02	93.96	101.07	96.83	93.31

Q7: Why did the 12 readings for 'apple' differ from each other?

- Summarise their response [i.e. "so you mean that the pupils did it differently and used different apples"] – is that what you meant? Are you happy with that explanation? [Try not to prompt for <u>more</u> causes of variation, but allow them to add more if they think of more until they're happy with their explanation]

Q8: Imagine if you'd been able to ensure that the pupils did *EXACTLY* the same thing as each other, would they have all got the same results as each other? Explain why you said that.

Probe 4:

A scientist wanted to see how temperature affected osmosis in a potato. She puts equal size chips in a 1 mol/dm^3 sugar solution at different temperatures and measured the change in mass of the 'chips' on a top pan balance and recorded the results as a percentage of the original mass after 4 hours. She repeated all the readings 3 times.

Percentage of original mass			
Temperature (ºC)	**1**	**2**	**3**
15	83.46	78.88	80.91
30	77.5	82.67	80.59
45	82.66	65.47	74.24
60	67.76	93.32	73.18

Q9: What does the data show? Explain all your reasoning.

- If the response is about the pattern in *columns* of data,
  - Ask why they're looking at that column.
  - Point to other columns and ask why the scientist repeated the readings.
- If the response is about *rows* of data,
  - Why might there be variation?
  - Ask what the scientist might conclude from all this data.

End of Interview 1

Annex 3.2 Questionnaire 1

on Pre-service Primary Teachers' handling of Data

Full name: _____ (Mr/Ms/Mdm)#

(as in matriculation card) #delete accordingly

Tutorial Group: _____

*Course code:

DCS100	ACS201

*Highest academic attainment:

A-levels	Polytechnic Diploma	Basic Degree	Post-graduate

*My highest level in Science was achieved at:

O-levels	A-levels	Polytechnic Diploma	Degree

*My Academic Subject in NIE is:

Physics	Biology	Mathematics	Other subjects

Please tick (√) to indicate your response

INSTRUCTIONS

This questionnaire is **not** for grading purposes.

Please respond to the questions **individually** without consultation with any resources including the Internet, colleagues, books, journals, etc.

Please answer as best as you can and do **not** leave any questions unanswered.

It is estimated that you would take about 40 minutes to complete all questions.

If you need any clarification or assistance, please call Mr Muhammad Shahrin at telephone no. **6790-3360** during office hours or email him at **shahrin.moorthy@nie.edu.sg**

Thank you for your assistance

Probe 1: Repeats

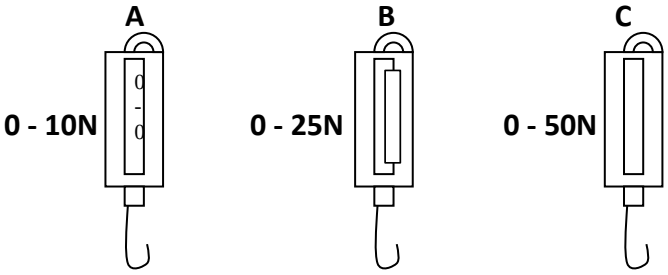
In science experiments, we often take more than one measurement or reading. The questions or statements in the table below look into your ideas concerning repeated measurements or readings.

Think carefully and **t*ick*** *(V)* the appropriate column to indicate your best response.

No.	Item	Agree	Disagree	Remarks, if any
1a	Two or three repeated measurements or readings are always enough. **1b.** If you disagree, propose how many times would be enough: _____.			
2	People who are good at doing experiments always get the same measurement or reading each time.			
3	You should go on taking measurements or readings until you know what the range of a variable is.			
4	You know you have got the right answer when you are able to get the same measurement or reading twice or more.			
5	***Ideally*** you should take as many measurements or readings as you possibly can.			
6	If you get one measurement or reading that is ***very*** different from all the others you should ignore it.			
7	Most measurements or readings when done several times would vary a bit no matter how careful you are.			
8	It is never possible to repeat a measurement or reading in exactly the same way.			
9	Precision means the values obtained in repeated measurements or readings are clustered closely together.			
10	The less an instrument's precision, the more is its uncertainty.			
11	A precise measurement may not necessarily be an accurate or 'true' measurement (and vice versa).			
12	We can perfect a measurement technique so that only one measurement will give a 'true' value.			
13	A fair test is one in which only the independent variable (whose values are changed constantly) has been allowed to affect the dependent variable.			
14	It is only fair to compare two similar experiments provided they have the same number of measurements or readings.			

Probe 2: Instrument

15 Which one is the **best** force meter to weigh an object of 7 Newtons?



Explain why you choose **A** or **B** or **C** or it does not matter.

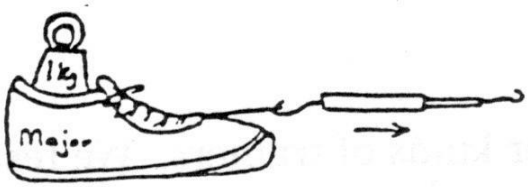
16 Which of the force meters, **A** or **B** or **C**, will you use to weigh an object of 18 Newtons?

Answer: _____

Explain your reason for using that particular force meter.

Probe 3: The Sole Test

Kumar, Ahmed and Lee did some work about how different surfaces affect the 'slippiness' of a running shoe by putting a 1 kilogram mass in the shoe and finding the amount of pull needed to drag the shoe along. They tested each surface twice.



Here are their results:

Type of surface	*Pull force (Newtons)*	
	1st trial	*2nd trial*
Soil on the school's playground	10	15
Grass on the school field	14	13
Carpet in the school's library	8	9
Cement floor in the school canteen	5	7

17 Why did they test each surface ***twice***?

18 They thought they had done everything the same but they did ***not*** get the same results. Suggest why.

19 Their teacher asked them which surface needed the ***most*** force to pull the shoe along. ***Tick*** (√) the one you most agree with:

- ☐ Kumar said it was the grass
- ☐ Ahmed said he couldn't tell
- ☐ Lee said it was the playground

Why did you choose that one?

Annex 3.3 Interview 2

Aim: to understand different conceptions about accuracy and precision, variation in repeated readings i.e. causes of variation in repeated readings; instruments and the quality of the data; students' tendency to put variation all down to 'human error'; distinguishing patterns in data with variation

Q1. A reading is described as being an 'accurate' value.

- Can you explain what do you understand by this statement?

  [*Check if students have the concept of accuracy being the closeness of a measurement to a true value*]

Q2. What do you understand by the term 'precision' as applied to repeated measurements?

  [*Check if students have the concept of precision as being the variation between repeated readings*]

Q3. A scientist reported that she obtained several melting point values for an organic substance and all are close to the value reported in the Data Booklet published by the Department of Science at the University.

- Describe the set of melting point values in terms of their accuracy and precision?

  [The values should be both accurate and precise]

Q4. Refer to statements I to IV below.

- Based on your understanding of 'accuracy' and 'precision', which describes the possible relationship between the two terms for a set of repeated readings?
  Think carefully. You may choose any combination as your answer.

I. Repeated readings can be both accurate and precise.
II. Repeated readings can be accurate but imprecise.
III. Repeated readings can be precise but inaccurate.
IV. Repeated readings can be both inaccurate and imprecise.

[All; I to IV]

Q5. If scientists use the same instrument several times to take repeated measurements of the same variable, would they get *exactly* the same reading?

- If the answer is no, then ask them to explain why?
  - If the response is about experimental 'error' or 'mistakes', then ask students to clarify their understanding of 'error' that causes the variation?

  [*Check if students were actually referring to human/instrument error*]
  - If the response includes any of these terms: 'random error' / 'uncertainty', ask them to clarify what do they mean?

  [*Check if students have the concept that there is always imprecision/variation between repeated readings*]

Q6. Examine the table below and answer the following question.

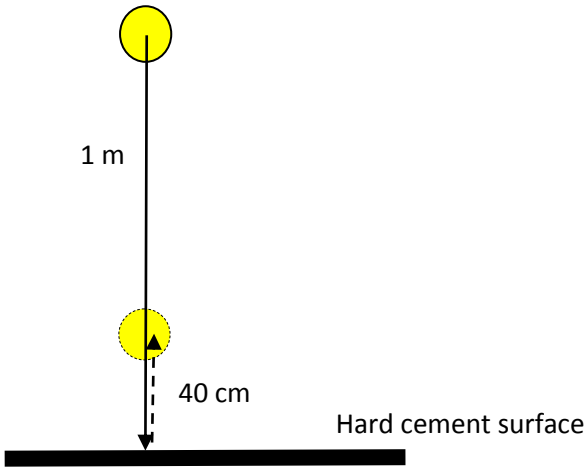
Melting ice investigation

Where 80g of ice was placed	**Volume of water in cm^3**
On the classroom floor	46
In the cabinet	39
On window sill	64
In refrigerator	10
Along the corridor	26

When fully melted, 80g of ice will be about $80cm^3$ of water. Of what capacity beaker would you use for the ice: $500cm^3$, $250cm^3$, $100cm^3$, or it does not matter?

Probe 7

Imagine you simply drop a rubber ball from the height of one metre and then measure its rebound height against the metre rule. It bounced back to 40 cm the first time.



Imagine you bounced it *9 more* times from the same height and each time taking note of the first bound height.

Q7. What would the 9 other first bound heights look like? Fill in imaginary results into the table below.

Your name: _____

Bounce number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Bounce height (cm)	**40**									

- If the data is varied,
  - Ask why is the data varied?
  - Ask if the differences between rebound heights (variation) become smaller/larger/either as they collected more data? Explain why?
  - Ask if they would continue to take more data after bounce number 10?
  - Ask when will they stop? Why stop at that bounce number?

Q8. If they were to report the rebound height of the rubber ball as a measure of its 'bounciness',
  - Ask how they would go about it?
  - Ask them to explain how they arrived at the value? Why that value?

[Check if students choose the 1st value, the average, median or mode value, or a range or spread of values]

Q9. Two students **A** and **B** carried out the same investigation with different rubber balls of the same brand and reported the results shown in the table below.

Bounce number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Student **A** Bounce height (cm)	40	45	36	50	39	42	42	42
Student **B** Bounce height (cm)	43	44	43	41	38	38		

Both students **A** and **B** reported bounce height of 42 cm by adding all the rebound heights and dividing by the total number of bounce for each.

- With whom, **A** or **B**, do you most closely agree? Explain your choice.

*[Check if students agree with **A** because more repeats/higher occurrence of '42', or **B** because smaller range/ student A had considered '50' probably an experimental error, and vice-versa]*

Q10. Instead of measuring the first bound height by eye against a metre rule, the students used a *self-automated* digital camera to record the first bound height.

- Would they obtain similar results (as varied as before)?
- Would they get more of the same results say more times of 40 cm? Explain why?
- Would they get a more accurate result for the average rebound height? Explain why?
- In your opinion, how many repeats should be appropriately conducted (as many, or more, or less than using the eye)?

Probe 8: The Pendulum

A student wanted to see how the length of a pendulum affects the period T, which is defined as the time taken for one complete swing. She took the time taken using a stopwatch for the pendulum to swing to-and-fro twenty times. She repeated the same measurement at different lengths three times. The results were recorded as shown in the table below.

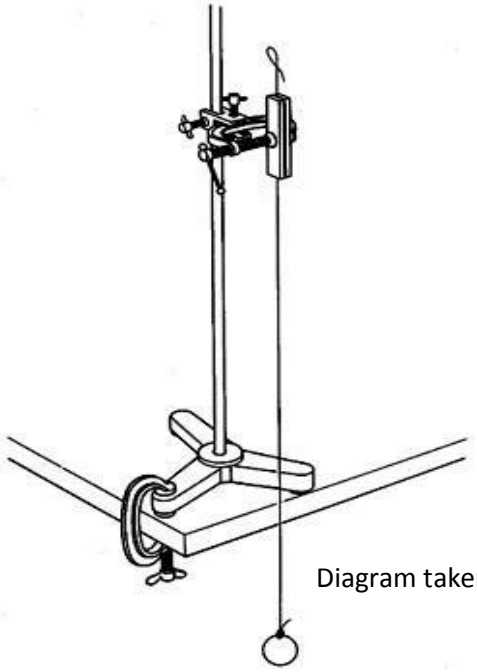


Diagram taken from www.practicalphysics.org

Length of pendulum in cm	Time take for 20 swings in seconds		
	1	2	3
40	26	25	28
60	31	32	32
80	36	36	41
100	40	47	41
120	50	43	44

Q11.Describe in detail what the results in the table show you.

- Are there concerns with any of the data shown in the table?
- If you were the student, how would you proceed with the experiment?

[C*heck if students were to suggest repeating experiment for the fourth time or at certain lengths only, or increase the range for length, etc*]

Probe 9: Osmosis

A student wanted to study how temperature affects osmosis in a potato. She placed 60-mm size chips in water at different temperatures and measured the change in length of the 'chips' after 4 hours. She repeated all the readings 3 times.

Temperature (°C)	Length of 'Chips' (mm)		
	1st trial	2nd trial	3rd trial
15	76	69	70
30	72	80	71
45	83	80	75
60	69	76	65

Q12. What does the data show? Explain all your reasoning.

- If the response is about the pattern in *columns* of data,
  - Ask why they're looking at that column.
  - Point to other columns and ask why the student repeated the measurements.
- If the response is about *rows* of data,
  - Why might there be variation?
  - Ask what the scientist might conclude from all this data.

Annex 3.4 Questionnaire 2

Full name: _____ (Mr/Ms)#
#delete accordingly

Tutorial Group: _____

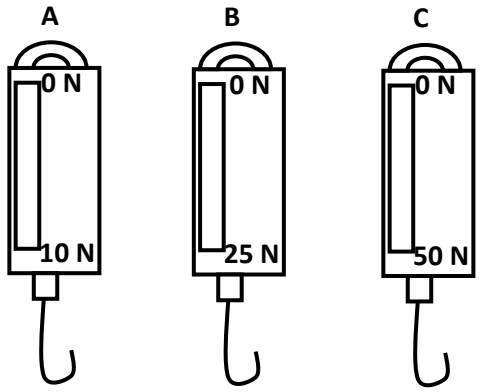
INSTRUCTIONS

1. This questionnaire is **not** for grading purposes.
2. Please respond to the questions **individually** as best as you can.
3. In some questions, you would be given more than one option, please **circle** the one option that represents your answer.
4. It is estimated that you would take about 40 minutes or less to complete all questions.

1 The Instruments Test

(a) **(i)** Which forcemeter would you choose to weigh an object of around **8 Newtons**?

Circle your answer	**It doesn't matter which**	**A**	**B**	**C**



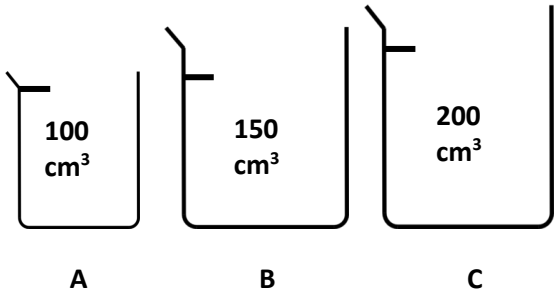
(ii) Explain your choice in (a) (i)_____

_____

_____

(b) **(i)** Which beaker would you choose to get a volume of water of around **80 cm³**?

Circle your answer	**It doesn't matter which**	**A**	**B**	**C**



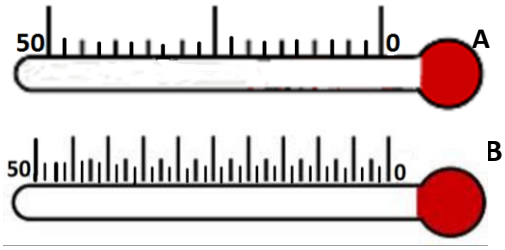
(ii) Explain your choice in (b) (i)_____

_____

_____

(c) **(i)** Which 50°C thermometer would you use to measure the room temperature accurately?

Circle your answer	**It doesn't matter which**	**A**	**B**



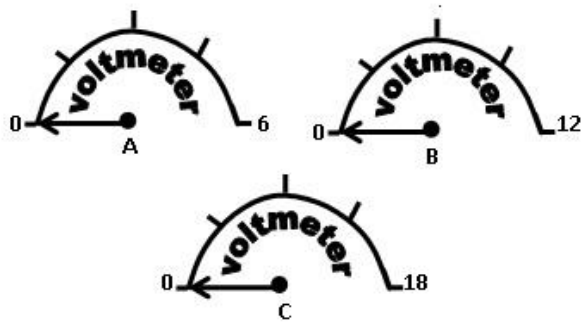
(ii) Explain your choice in (c) (i) _____

_____

_____

(d) **(i)** Which voltmeter would you choose to measure a 5 volts dry cell?

Circle your answer	**It doesn't matter which**	**A**	**B**	**C**



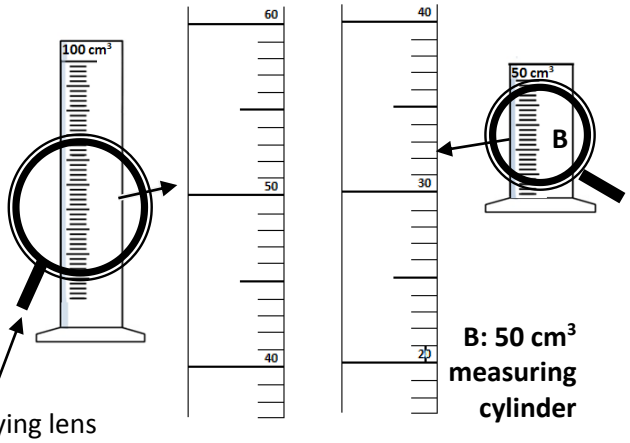
(ii) Explain your choice in (d) (i)_____

_____

_____

(e) **(i)** Which measuring cylinder would you use to measure 35 cm^3 of salt solution?

Circle your answer	**It doesn't matter which**	**A**	**B**



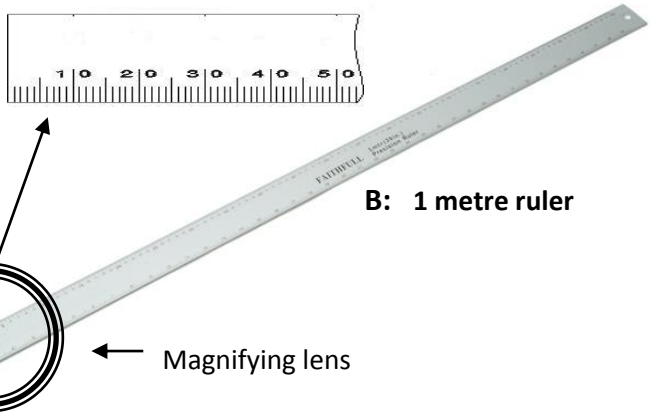
(ii) Explain your choice in (e) (i) _____

_____

_____

(f) **(i)** Which ruler would be able to measure an 8.0 cm length of string accurately?

A: 10 cm ruler



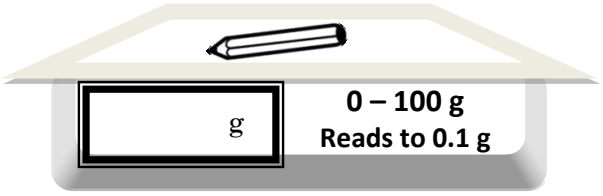
Circle your answer	**It doesn't matter which**	**A**	**B**

(ii) Explain your choice in (f) (i) _____

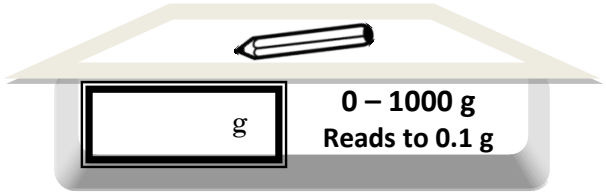
_____

_____

(g) **(i)** Which digital weighing balance would be able to measure the weight of a pencil more accurately?



A



B

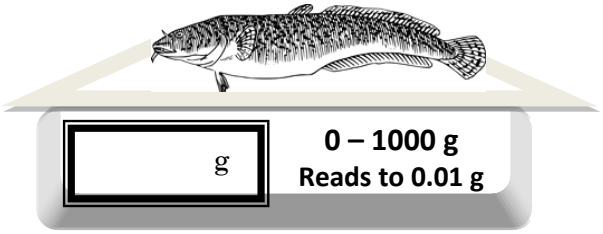
Circle your answer	**It doesn't matter which**	**A**	**B**

(ii) Explain your choice in (g) (i) _____

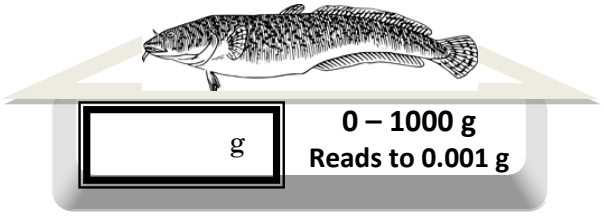
_____

_____

(h) **(i)** Which digital weighing balance would be able to measure the weight of the fish best?



A



B

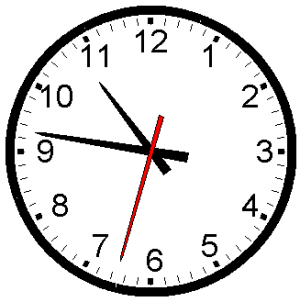
Circle your answer	**It doesn't matter which**	**A**	**B**

(ii) Explain your choice in (h) (i)_____

_____

_____

(i) **(i)** Which clock would be able to tell you the time best?



A

B

Circle your answer	**It doesn't matter which**	**A**	**B**

(ii) Explain your choice in (i) (i)_____

_____

_____

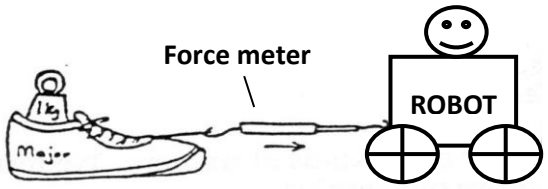
2 Repeats

In an investigation, we often take repeated readings for each value of the independent variable. The statements in the table below are about repeated readings. Think carefully and give your ***best*** response i.e. whether you agree or disagree with the statement.

	Statement	AGREE	DISAGREE
(a)	Three repeated readings are all we need.	**A**	**D**
(b)	Most readings when done several times will vary a bit no matter how careful you are.	**A**	**D**
(c)	We decide the number of repeats after we have done a few readings.	**A**	**D**
(d)	If you get one reading that is very different from all others you should leave it out of your calculations.	**A**	**D**
(e)	People who are good at doing experiments always get the same reading each time when making a measurement.	**A**	**D**
(f)	The variations in repeated readings are due to human errors only.	**A**	**D**

3 The Sole Test

Here are the results of an investigation into how different surfaces affect the 'slippiness' of a trainer. The experiment is carried out by putting a kilogram mass in the shoe and finding the amount of pull needed to drag the shoe along. A remote controlled robot was used for this purpose. Each surface was tested several times and the results for 3 trials are shown below.



Type of Surface	*Pull force (Newtons)*		
	1	*2*	*3*
Soil on the school's playground	10	10	16
Grass on the school field	10	12	14
Carpet in the school's library	4	6	6
Wooden floor in the school hall	7	6	3

(a) Why do you think each surface was tested ***more than once***? Circle your answer.

A to check the first reading
B to see how much the readings vary
C to get the same reading a few times
D to practice and get better

For each of the following questions, choose your answer by circling one of the options, and then explain why you chose that one in the corresponding box.

(b) Looking at the table of readings, answer the following questions.

(i) Which surface needed the ***most*** force to pull the shoe along?

A the grass
B the playground
C I cannot tell which

(b)(ii) Why? _____

_____

_____

(iii) Which surface needed the ***least*** force to pull the shoe along?

A the carpet

B the wooden floor

C I cannot tell which

(b)(iv) Why? _____

_____

_____

(c) A similar investigation on tiled floor surfaces found the results in the table below.

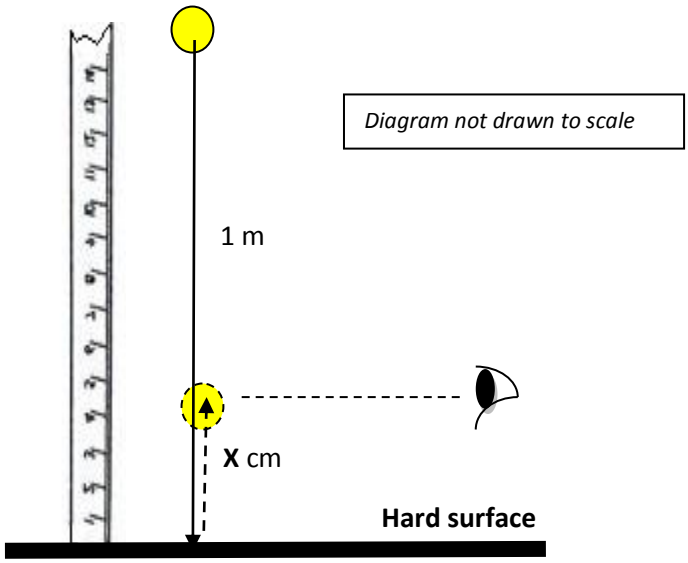
Location of the tiled floor surface:	***Pull force (Newtons)***						
	1	***2***	***3***	***4***	***5***	***6***	***7***
(I) Classroom	14	11	12	13	10		
(II) HOD's Office	10	11	13	13	13		
(III) Teachers' Common Room	10	11	12	13	14		
(IV) Science Laboratory	11	16	11	12	10		
(V) School Canteen	10	14	10	14	12	14	10

For each of the following comparisons, choose the surface that needed ***more*** force to pull the shoe along? Circle your answer and explain why.

No.	**Comparison**	**Circle your answer**			**Why you choose that one?**
(i)	Between (I) Classroom and (II) HOD's Office	**(I)**	**(II)**	I can't tell which one	
(ii)	Between (I) Classroom and (III) Teachers' Common Room	**(I)**	**(III)**	I can't tell which one	
(iii)	Between (I) Classroom and (IV) Science Laboratory	**(I)**	**(IV)**	I can't tell which one	
(iv)	Between (I) Classroom and (V) School Canteen	**(I)**	**(V)**	I can't tell which one	
(v)	Between (II) HOD's office and (III) Teachers' Common Room	**(II)**	**(III)**	I can't tell which one	
(vi)	Between (II) HOD's office and (IV) Science Laboratory	**(II)**	**(IV)**	I can't tell which one	
(vii)	Between (II) HOD's office and (V) School Canteen	**(II)**	**(V)**	I can't tell which one	
(viii)	Between (III) Teachers' Common Room and (IV) Science Laboratory	**(III)**	**(IV)**	I can't tell which one	
(ix)	Between (III) Teachers' Common Room and (V) School Canteen	**(III)**	**(V)**	I can't tell which one	
(x)	Between (IV) Science Laboratory and (V) School Canteen	**(IV)**	**(V)**	I can't tell which one	

4 The Bouncing Rubber Ball Test

Jill conducted an experiment to find the bounciest ball; she compared 3 pairs of balls and took 20 readings for all of them. Each time, she would release the ball from a height of one metre and then measured its rebound height in centimetres against a metre rule (see figure below). For easy reference, the balls are referred to as **A**, **B**, **C**, **D**, **E** and **F**.



Study the tables below. If you had done the experiment, tell us ***how many repeats*** you would do to find out which ball was **bouncier**? Circle your answer and explain why you choose that one.

(a) Comparing Ball A and Ball B

Bounce number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ball A	29	29	27	28	29	28	30	26	28	30	25	28	26	27	31	31	30	29	30	28
Ball B	28	30	29	28	29	31	29	27	28	27	28	25	31	26	29	30	29	25	28	27

1 about 3

2 about 10

3 about 20

Because _____

_____

_____

(b) Comparing Ball C and Ball D

Bounce number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ball C	20	21	23	20	21	19	20	23	21	20	19	21	19	17	20	19	22	20	20	21
Ball D	44	43	42	39	40	41	39	40	43	43	39	42	39	40	39	43	38	39	37	43

1 about 3

2 about 10

3 about 20

Because _____

_____

_____

(c) Comparing Ball E and Ball F

Bounce number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Ball E	22	25	27	25	20	14	25	18	20	16	22	20	16	14	22	15	18	11	19	21
Ball F	26	30	26	31	35	28	25	37	32	36	30	36	34	31	35	36	29	35	38	40

1 about 3

2 about 10

3 about 20

Because _____

_____

_____

5 (a) Starting an Investigation

(i) You wish to see how independent variable **X,** which has continuous values**,** affects dependent variable **Y** in an investigation. Which plan shows the sequence you would take your first 4 readings? Circle your answer and explain why?

A

Independent variable **X**	Dependent variable **Y**		
↓	**(1)**	**(4)**	
	(2)		
	(3)		

B

Independent variable **X**	Dependent variable **Y**		
↓	**(1)**	**(2)**	**(3)**
	(4)		

C

Independent variable **X**	Dependent variable **Y**		
↓	**(1)**	**(2)**	**(3)**
	(4)		

(ii) Explain why you chose **A, B** or **C**?

_____

_____

_____

5 (b) What next in an Investigation

Each of the tables below shows the first **9 values** you have recorded for a dependent variable **Y** responding to changes of an independent variable **X** in separate experiments. Answer the corresponding question. For the multiple-choice questions, circle your answer and then explain why?

(i)

independent variable X	Dependent variable Y				
	1	**2**	**3**	**4**	**5**
	F				
20	*92*	*72*	*84*	**A**	**B**
30	*86*	*82*	*80*	**C**	
40	*68*	*70*	*80*		
	D				
	E				

(I) What would be your next **2** readings?

(1) **A** & **B** (2) **A** & **C** (3) **D** & **E** (4) **E** & **F**

(II) Why? _____

_____

_____

(ii)

independent variable X	Dependent variable Y				
	1	**2**	**3**	**4**	**5**
	F				
20	*90*	*81*	*77*	**A**	**B**
30	*73*	*80*	*80*	**C**	
40	*72*	*79*	*81*		
	D				
	E				

(I) What would be your next **2** readings?

(2) **A** & **B** (2) **A** & **C** (3) **D** & **E** (4) **E** & **F**

(II) Why? _____

_____

_____

(iii)

independent variable X	Dependent variable Y				
	1	**2**	**3**	**4**	**5**
	F				
20	*80*	*70*	*58*	**A**	**B**
30	*80*	*78*	*75*	**C**	
40	*80*	*86*	*88*		
	D				
	E				

(I) What would be your next **2** readings?

(3) **A** & **B** (2) **A** & **C** (3) **D** & **E** (4) **E** & **F**

(II) Why? _____

_____

_____

(iv)

independent variable X	Dependent variable Y				
	1	**2**	**3**	**4**	**5**
10	A	B	C	D	E
20	***80***	***81***	***82***	F	G
30	***82***	***89***	***83***	H	I
40	***83***	***84***	***85***	J	K
50	L	M	N	O	P
60	Q	R	S	T	U

(I) What do the data (in numbers) show?

_____

_____

_____

(II) Which letters can represent your next **2** readings and why?

____ &____ because_____

_____

(v)

independent variable X	Dependent variable Y				
	1	**2**	**3**	**4**	**5**
10	A	B	C	D	E
20	***82***	***82***	***83***	F	G
30	***84***	***82***	***81***	H	I
40	***84***	***83***	***80***	J	K
50	L	M	N	O	P
60	Q	R	S	T	U

(I) What do the data (in numbers) show?

_____

_____

_____

(II) Which letters can represent your next **2** readings and why?

____ &____ because_____

_____

End of Questionnaire

Thank you

Annex 3.5 Sample of Invitation letter for validation

To:

_____

NSSE Academic Group
National Institute of Education

Dear Colleague,

REQUEST FOR ASSISTANCE IN VALIDATING QUESTIONNAIRE

I am currently working on my EdD thesis with Durham University under the supervision of Ros Roberts and Richard Gott. I am collecting data on pre-service primary science teachers' handling of experimental data.

Specifically, the aims are to understand different conceptions about accuracy and precision, variation in repeated readings i.e. causes of variation in repeated readings, instruments and the quality of the data, students' tendency to put variation all down to 'human error', distinguishing patterns in data with variation, and using tables and graphs.

This questionnaire will serve to corroborate and substantiate findings from two rounds of interviews conducted earlier with 55 pre-service primary school teachers. This questionnaire is also been piloted with 37 individuals from the same group.

For your information, the questions have been taken from the research group in Durham University led by the team of Richard Gott and they have largely been validated for their use in the UK. My intent is merely to validate the questionnaire for the local context.

Your assistance is greatly appreciated.

I would be happy if you could go through the questions and provide me feedback _____ using the attached checklist or you can annotate your comments on the questions themselves.

Thank you and I am grateful for your effort.

Yours sincerely

Md Shahrin
H/p 98509353; Mailbox 23; Room 109E

CHECKLIST

No.	Questions	Agree	Disagree	Remarks (if any)
1	The instructions for the questions are clear.			
2	The questions are worded clearly and can be understood by pre-service teachers.			
3	Pre-service teachers should be able to understand the language used in the questions.			
4	The terms and phrases peculiar to science investigations deployed in the questions will not pose any difficulty to the respondents.			
5	The questions are not ambiguous.			
6	The questions are in line with the knowledge and skills expected of pre-service primary science teachers.			

Please write if you have other comments/feedback:

_____

_____

_____

_____

Name of Validator: _____

Annex 3.6 Samples of Validators' Feedback to Questionnaire 1

CHECKLIST

No.	Questions	Agree	Disagree	Remarks (if any)
1	The instructions for the questions are clear.	✓		
2	The questions are worded clearly and can be understood by pre-service teachers.	✓	✓	A few questions - there are ambiguities as indicated.
3	Pre-service teachers should be able to understand the language used in the questions.	✓	✓	Not all - comments indicated on worksheet
4	The terms and phrases peculiar to science investigations deployed in the questions will not pose any difficulty to the respondents.	✓		How about categoric/continuous variables - would students know what this means?
5	The questions are not ambiguous.	✓	✓	A few questions have ambiguities as indicated.
6	The questions are in line with the knowledge and skills expected of pre-service primary science teachers.	✓		

Please write if you have other comments/feedback:

Overall, it is a good instrument. Some refinements needed to improve clarity before it can be used on pre-service teachers. I would be pleased to discuss this with you if necessary.

Name of Validator: R. Subramaniam

11/4/2010

CHECKLIST

No.	Questions	Agree	Disagree	Remarks (if any)
1	The instructions for the questions are clear.	✓		Generally, instructions are clear. I've suggested one change on line 2 of instructions to improve comprehension. You may want to relook at line 4, sounds a little old but can't think of any good change.
2	The questions are worded clearly and can be understood by pre-service teachers.		✓	Pls see comments given.
3	Pre-service teachers should be able to understand the language used in the questions.		✓	Pls see comments given.
4	The terms and phrases peculiar to science investigations deployed in the questions will not pose any difficulty to the respondents.	✓		Generally ok with sc. language but pls see comments given for some of them
5	The questions are not ambiguous.		✓	Pls see comments given.
6	The questions are in line with the knowledge and skills expected of pre-service primary science teachers.			Depends when they do this. Many pre-service Pri sc. teachers do not have science background. If these have not been covered they might have some diff.

Please write if you have other comments/feedback:

Language & clarity of questions items need to be improved. Minor areas would be formatting.

Name of Validator: Jennifer Yeo

1. Because I feel that when you're doing an experiment right, there are many variables so if you just depend on one reading the answer may not be accurate because if you do repeated readings certain things may change so you will get an average of every reading. Ermm…I'm not sure whether this is the right way to answer.

   You talk about accurate readings right so you do repeat reading to…

   Yes but maybe about 2, 3 times if I'm doing the experiment.

   So you do repeat readings to get an accurate value. Do you think scientist have this idea of some true value out there where they are trying to achieve?

   I think so but maybe not in all cases because before they start an experiment they should have a clear idea about what it is because they should have read up on this before starting the experiment, so they would have some things to follow so they would be pretty sure what the ideal reading should be or maybe if they are just starting a new experiment then maybe they may not have an idea. Depends actually.

2. Somewhere near…but I think it actually depends on the impact that they throw also

   Ok you're not throwing; you're just dropping it from one metre.

   Maybe about 50, 40 also, around there. Around the same

   Why do you think you'll get about the same reading?

   Because actually I feel it depends on the impact.

   Ok but you're not throwing, you're just dropping it.

   Just dropping ah? Oh, because you're dropping from the same height, you're using the same ball there's no change in much of the things used so it should be around the same, there's not much changes.

   Can you just write down the value for the second time you drop the ball?

   Maybe 45?

   So it's not exactly 40, its like 5cm.

Ya.

Can it go the other way, I mean less than 40?

Possible right but maybe not a lot.

Why do you think so?

Because I think it's plus minus, it's around that range

Ok plus and minus what?

Maybe not more than 5. But I think the chances it will be higher

Oh chances are higher?

Ya, because if I drop…I think…

OK.

Is it correct?

I, well there's no right answer over here, it doesn't matter, what comes to your mind, you just say it.

3. Maybe 41, 43, 45, 48, 42 that kind around that range

So everything is above 40 is it?

Ya to me I think it will be above 40.

Can you just explain again why it is not below 40?

Ok you said it is just dropped right from a height of 1m then you measure the rebound height against a metre ruler. Ok I feel… actually I don't really know how to explain this, based on my intuition and everything I think it is unlikely to fall below 40, although it may be possible but if it does I don't think it will be more than 5cm below 40 also, because 1m is actually quite high as well, and the ball is quite light.

4. **So let's say if you were to do the same action 50 times or 100 times, how many times will be above 40 or would you get…?**

   1m is not that tall ah so maybe there is a possibility that it will fall below 40 also, ya, 1m isn't that tall

   Ok what if now you were to do it 50 or 100 times, what sort of readings would you get?

   Ok maybe initially it will be more than 40, then slowly below 40.

   Ok why would it go below 40?

   Maybe because the person dropping gets tired the he releases less energy.

   So it has something to do with the person who does the experiment is it?

   Ya.

   Ok so what you said was if you do it 50 to 100 times it's a range.

   Yes, but not much difference.

   I think what you're alluding to is what we say a human error. Would it less error with more results?

   I think there would be a time between the 50 and 100 where the readings are around the same because the person is used to it already so it's easier to predict.

5. If I used a video to freeze the point, as in I predict the height based on the video, is it?

   No, if you look at the results form the video would it be the same as the ones you had in the 10 in the table?

   Actually maybe the video will have a higher reading.

   A higher reading means what?

   As in maybe a higher point in the instant in which the ball reaches the highest. Because if you throw something the video can capture it at the exact second but if you're using your naked eyes to look at the metre ruler may be some ambiguous reading, like plus minus.

Oh ok. So you felt that the video reading will always give you a higher reading than the eye reading?

Actually not the case because you're using your eyes to take the video so even if you are very accurate, you need to look then press the button, so I feel that the metre ruler will be the more accurate one because you're not going through two things.

6. Personally I don't think so but I think there's a lot of instrument that can be comparable to perfect but depends on what you use to measure, I mean what you're measuring.

A perfect instrument – what does it mean to you?

Something that…

Is it like error-free kind of thing?

Maybe and then it takes into consideration all the variables so it will be a perfect one.

So assuming I have such an instrument, a hypothetical case in your case do you think it will give me the same reading over and over again?

A perfect instrument…no.

What may cause the difference then?

Ok let's say if you want to use an instrument to measure weather. Weather you have a lot of different things to consider like the sun, water, the atmosphere so even I you have a perfect instrument that takes into consideration all these changes in variables, then readings will still be different because if I'm using weather everything is constantly changing. I feel that way.

7. I cannot remember what osmosis is. Osmosis is the water, the plant changes…

Remember the movement of water from a higher potential to a lower potential across a semi-permeable membrane. Can you recall?

Roughly; is it something to do with the plant cycle?

Not really the plant cycle. Can you describe the experiment you did in primary school?

I cannot really remember. It involves water, leaves. I think the osmosis we learned was something to do with stomata in the same chapter.

So you learned about the applications of osmosis. I do not know if you have done this experiment where you put equal size potatoes into water and then when you look at the size again it has sort of expanded in its length because water has gone into the potatoes so this is a living process. What do you think could have happened here? Why do you think the pupils show these results?

Why did the 12 readings of apple differ is it? It's the same apple right.

You think that could be the reason – the apple?

Maybe I'm not really sure. Maybe the different times it absorbs the sugar solution will result in a change in mass, as in the different rate of absorption.

Ok so you're implying the apples are different is it?

The absorption rate of the apple chip.

Maybe because you have not done this so you find it a bit difficult to picture it in your head.

8. Based on other experiments because no two experiments got the same results it's very unlikely. From my experience all the experiments in like physics, chemistry, there's no two experiments as my partner that we got the exact same results; they'll be a little difference, maybe not much but I don't recall any same results.

What can contribute to the difference?

Maybe as an individual how you operate the experiment is one factor also then the material also; different variables will affect the results.

Ok it depends on the human and the materials, is it?

Ya.

9. The data shows me that at different temperature, the percentage of original mass is different. And each size chip actually differs from each other even at the same temperature.

Can you show me what are you looking at to tell me this?

Ok the first one is at different temperatures, the data are different [*the percentage of original mass*]. Then the second thing you can tell is even at the same temperature, the equal size chip 1, 2 and 3 they have different percentage mass.

Why is there this difference, we call it variation right, for a particular temperature you are looking at 1, 2 and 3 and you see that they are different. Why is there this difference or this variation? Maybe you can hypothesize why there is a difference?

Is it possible three of the chips in the same sugar solution or each one in one mole, one mole, one mole, or, all three in one mole? Maybe this could be a reason.

So if I have all 3 chips in one Petri dish is it?

Ya that's why there may be a change.

So if all 3 chips are in the same Petri dish there might be a difference?

Because the absorption may be different since it is only one mole of sugar solution. Then if it is in different there may be also a change because she recorded it after 4 hours so maybe time is a factor also.

Ok but this is exactly at 4 hours you know when she looked at the masses. So that would affect?

Maybe each chip absorption rate is different.

Ok. Now if you look at this ah, if you are the scientist what can you conclude from the data?

Oh I was going to say that at which temperature the absorption rate will be higher. At 60, piece 2 is actually higher, at 15, piece 1 is actually higher, so it is a bit hard to conclude. Maybe I'll find the average of all the readings divide by 3 then I'll realise at which temperature which percentage of original mass is higher. Plus all 3 then divide by 3 then write the readings at the side.

So do you think taking an average is important in this case?

Yes I think more important because you have different readings.

Ok so in all experiments that you have done is that usually the case where you repeat readings to take the mean? Is that the case? Or is it sometimes you just do one then you're happy with it?

I think that's the case, especially for physics but I can't really remember for chemistry.

1. Means the systematic error and all the errors are proven notwithstanding and the answers presented are as good as you can get and to the accuracy you specified. You must be able to see the tell-tale signs and how you arrive at the accurate value to see whether it's really accurate.

   You mentioned systematic error; can you briefly explain its meaning?

   To the best of my knowledge, systematic error is like the process itself there is a mistake like in the interpretation and then after that, you do the things wrongly. Your readings and your experimental technique may be correct but because of the error in judgement, you use the wrong method to arrive at the answer. The other is random error. Random error cannot be eliminated perfectly but systematic error can be avoided.

2. Like I just said when you do repeated readings, it may reduce random error but it doesn't eradicate them. For example, human lag time is 0.3 seconds so no matter how precise an instrument can be in measuring time, it doesn't make sense to estimate it in more than 1 decimal place. When you look at the precision of an answer, you take into consideration all the factors like how you conduct an experiment.

   Do these two terms appear the same to you?

   No, they are different to me. Accuracy is like you are sure that the system you carry out or the process is good, and the method you do is the correct one, so you arrive at an accurate answer. An accurate answer may not be as precise as a wrong answer. The wrong answer can be very precise but if systematic error occurs, and it is repeated in the measurements a lot of times, they may arrive at a precise answer but the answer is still wrong. So it can be precise but wrong.

3. When you do an experiment, there is an experimental and theoretical yield. So you have to look at the data booklet to see whether they are giving you the theoretical or experimental yield. If you say your answer is close to it then you have to see whether it is above or below it. Because by logical reasoning, you cannot achieve something that is more than the theoretical yield. In that case, something must be really wrong; it could be a systematic error or the experimental techniques that you used. But if you say that it's slightly under, it might then be the environmental effect or your experimental technique. To get the theoretical yield is impossible but a 90% yield will be considered good under normal conditions.

   What about in terms of melting points?

   For the melting point, you have to look at how pure the substance is. If the substance is pure, the melting point will be accurate, if it's impure, the melting point might vary. During your experiment, you might have contaminated the substance.

In terms of precision, you've to look at the apparatus you are using to measure the melting points. If you used a lab thermometer, it will only go up to 0.5^{o}C but if you used something more precise, you may get something like 0.01^{o}C. I am talking about calibration here

4. I choose III but repeated readings can also be both accurate and precise. Actually all 4 scenarios can be possible; if you're talking about the correct technique, I feel I and III make sense but if you're talking about the students not having very good lab skills then repeated readings don't lead to precise readings.

5. It will depend on the instruments they used; if they use those ancient tools like a weighing balance, they might not. But if they do digitally, they would be able to get back the same reading.

6. (a) What is this? [Pointing to the volume of water in cm^3].

These are the volumes of water at different locations.

The 100 cm^3 one.

Why?

If the readings are taken off the beaker and if it's part of the experiment, then I'll use the 100 cm^3 one because it's much more easier to read off from the 100 cm^3 as its calibration is much more precise than the 250 cm^3 or the 500 cm^3. The rise of per cm^3 of water per cm is much easier to see.

7. The data have slight variations.

Why?

The instrument that is used is a metre rule and your eyes. The metre rule calibration is by centimetre. It did not specify where to catch the height. To catch the highest point, the eyes must do a thorough job to catch it. But most of the time, human errors exist and you never know whether it was the highest point.

Would the difference between the heights become smaller, bigger or the same with more data?

The variance should remain the same with more times, not considering that you may be tired after some time.

Would you continue to take more data after bounce number 10?

If you want a more precise answer, you'll have to do more.

When will you stop?

When you look at the variance of the data, you're probably looking at about 5 to 10 cm so if you do about 5 to 8 times, and you can average it out pretty well already. Doing another 10 is more than enough because if your variance is 5 cm, and if you do it 20 times, you have actually minimised it and get a pretty accurate mean. So, it depends on the variance, I'll do 5 to 8 times and check the variance, I'll stop if it's small but I'll do another 10 if it's big.

8. I'll take the mean of all 10.

9. Just by looking at this. Student A uses the mode of the data but I'll still go with B

Why?

Because his readings are more consistent, they don't vary that much from 42. The biggest variance here is 4 whereas for student A, he has a result of 50 which is 8 cm more than the mean, and at bounce number 3, he has a 36, which is 6 cm away from 42. If you plot the graph, the best fit line will not include some of the points which show he has some experimental errors showing his experimental techniques or his reading is a bit off. For student B, although he has fewer readings, his readings were more consistent, so it gives me good reason that his answer is accurate.

Did you take into consideration that student A has 42 cm in bounce number 6, 7 and 8?

That's right. I did not ignore them. You can see that the reading of 1, 6, 7 and 8 were quite consistent. But there were some off readings that were taken. It cast some doubt on me whether student A has conducted the experiment in the correct manner. If he didn't and he arrived at 42 several times, this 42 cannot be taken as an accurate value.

10. **Would they obtain similar results as before?**

Result-wise it should be the same because the metre rule's calibration is 0.1 cm. I would expect the digital camera would be able to give an accuracy of 0.1. So, I would still say it would be close to 42, there'll be slight variations but less than using the eye alone.

Would they get more of the same results like 40 cm?

Yeah, I would think so.

Would they get a more accurate average rebound height with the digital camera?

Yeah, I think so.

Would you still do 10 times? When do you stop?

No, for the digital camera, six readings would be enough. This falls between 5 to 8 times. This would be less than using the eye alone.

11. It shows the length of the line and the time taken for 20 swings. For every single length, she took readings. So in essence, she wants to find out how the length of the pendulum affects the period of the oscillation; the period of the oscillation increases as the length increases.

Have you any concern with the data shown in the table?

Maybe the time taken for the 80 cm is longer than the one taken at 100 cm. The general trend is that as the length increases, the time taken for 20 swings also increases, so this must be an abnormality. Also the accuracy of the readings, the stopwatch should give me 0.1 seconds accuracy.

How would you proceed with the experiment?

If I spot an abnormality, I'll do another set of 3 for that particular length again. I'll compare my results to see if 41 is a one-off error. I use the new set of readings as my accurate and I'll replace the first set with the second set of readings. After that I'll tabulate my means and I'll plot a line graph; the length will be the x-axis and the time taken will be the y-axis.

12. The data show the length of the chips increases as the temperature where the osmosis takes place increases. That's the general trend. The trend bucks at 60^{o}C where instead of increasing above 83, it goes below 70^{o}C, which is below the reading of 15^{o}C.

Why did the student repeat a second and a third time?

To achieve better reading; she was trying to check whether the readings are consistent. If they vary too much, she knows there is an error and she has to look at how she conducted the experiment.

Annex 3.9 Sample of completed Questionnaire 1 (P1Q1)

Repeats

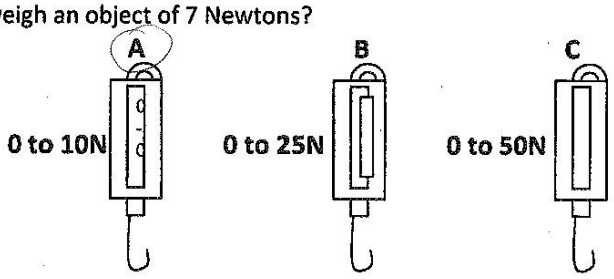
In science experiments, we often take more than one measurement or reading. The questions or statements in the table below look into your ideas concerning repeated measurements or readings.

Think carefully and ***tick*** *(√)* the appropriate column to indicate your best response.

No.	Item	Agree	Disagree	Remarks, if any
1	Two or three repeated measurements or readings are always enough. If you disagree, propose how many times would be enough: _____.		✓	- Depends if there is inconsistent readings - Depends on reasonable degree of error expected. - Usually more is better.
2	People who are good at doing experiments always get the same measurement or reading each time.		✓	Not always. often maybe. once again depends on nature of experiment.
3	You should go on taking measurements or readings until you know what the range of a variable is.	✓		
4	You know you have got the right answer when you are able to get the same measurement or reading twice or more.		✓	Maybe ~~two~~ thrice or more.
5	***Ideally*** you should take as many measurements or readings as you possibly can.	✓		
6	If you get one measurement or reading that is ***very*** different from all the others you should ignore it.		✓	- Ready should ~~not~~ be accounted for in report as experimental errors
7	Most measurements or readings when done several times would vary a bit no matter how careful you are.	✓		
8	It is never possible to repeat a measurement or reading in exactly the same way.		✓	
9	Precision means the values obtained in repeated measurements or readings are clustered closely together.		✓	~~Accuracy~~. Sounds more like accuracy ??
10	The less an instrument's precision, the more is its uncertainty.	✓		
11	A precise measurement may not necessarily be an accurate or 'true' measurement (and vice versa).	✓		other variables might cause inaccuracy even though precise.
12	We can perfect a measurement technique so that only one measurement will give an accurate or 'true' value.			Not sure.
13	A fair test is one in which the dependent variable responds only to an independent variable, which is often changed by a certain fixed amount in an investigation.	✓		Should be repeated too.
14	It is only fair to compare two similar experiments provided they have the same number of measurements or readings.	✓		

Instrument

15 Which one is the **best** force meter to weigh an object of 7 Newtons?



Explain why you choose **A** or **B** or **C** or it does not matter.

A. It should be more precise. Same reason as use 10 ml measuring ~~cylinder~~ cylinder to measure 10 ml solution than ½ of 20 ml measuring cylinder.

16 Which of the force meters, **A** or **B** or **C**, will you use to weigh an object of 18 Newtons?

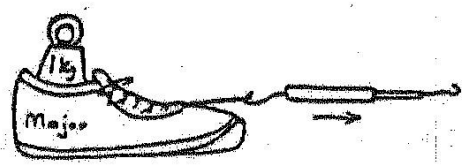
Answer: B

Give your reason for using that particular force meter.

~~It~~ A ~~is too~~ does not accomodate 18 N in its range. B is ~~the~~ next must ~~pres~~ precise in measurement.

The Sole Test

Kumar, Ahmed and Lee did some work about how different surfaces affect the 'slippiness' of a running shoe by putting a 1 kilogram mass in the shoe and finding the amount of pull needed to drag the shoe along. They tested each surface twice in the same area at all the different locations.



Here are their results:

Type of surface	*Pull force (Newtons)*	
	1st trial	*2nd trial*
Soil on the school's playground	10	15
Grass on the school field	14	13
Carpet in the school's library	8	9
Cement floor in the school canteen	5	7

17 Why did they test each surface ***twice***?

For greater accuracy.

18 They thought they had done everything the same but they did ***not*** get the same results. Suggest why.

There could have been other conditions affecting the ~~reads~~ results. Also, it could be caused by a acceptable range for error.

19 Their teacher asked them which surface needed the ***most*** force to pull the shoe along. ***Tick*** (V) the one you most agree with:

- ☐ Kumar said it was the grass
- ☐ Ahmed said he couldn't tell
- ☐ Lee said it was the playground

Why did you choose that one?

The range of the results for the grass & playground are not defined from each other & the range overlaps (fall under common range).

Annex 3.10 Sample of completed Questionnaire 2 (P2Q2)

Full name: _____ (Mr/Ms)

Age: 22 Course Title: ~~DIT~~ DCS Highest level Pass: A-Level/Poly Diploma/Degree#

#delete accordingly

INSTRUCTIONS

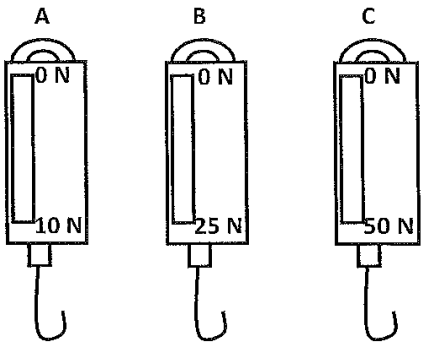
1. This questionnaire is **not** for grading purposes.
2. Please respond to the questions **individually** as best as you can.
3. In some questions, you would be given more than one option, please **circle** the one option that represents your answer.
4. It is estimated that you would take about 40 minutes or less to complete all questions.

---

1 The Instruments Test

(a) **(i)** Which force meter would you choose to weigh an object of around **8 Newtons**?

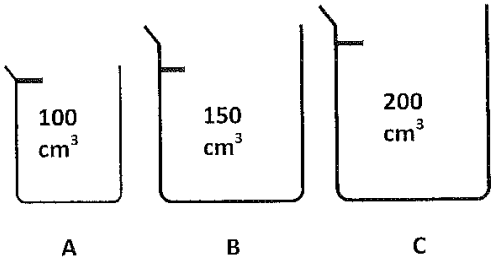
Circle your answer	It doesn't matter which	A	B	C



(ii) Explain your choice in (a) (i) 10N is closest to 8N.

(b) **(i)** Which beaker would you choose to get a volume of water of around **80 cm^3**?

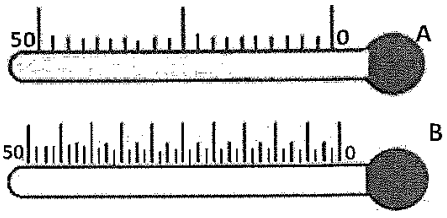
Circle your answer	It doesn't matter which	A	B	C



(ii) Explain your choice in (b) (i) 100cm^3 is closest to what we are finding - 80 cm^3

(c) **(i)** Which 50°C thermometer would you use to measure the room temperature accurately?

Circle your answer	~~It doesn't matter which~~	A	B

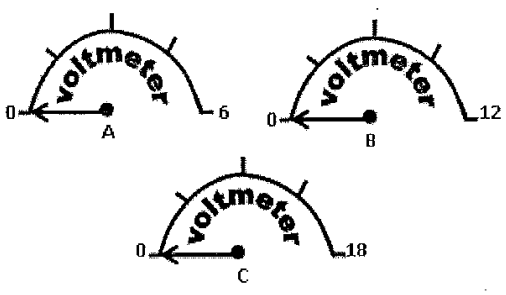


(ii) Explain your choice in (c) (i) 'B' is chosen in order to be exact when measuring temperature Eg → 39.7°C

(d) **(i)** Which voltmeter would you choose to measure a 5 volts dry cell?

Circle your answer

It doesn't matter which	A	B	C



(ii) Explain your choice in (d) (i) Enable us to see the measurement clearly accurately

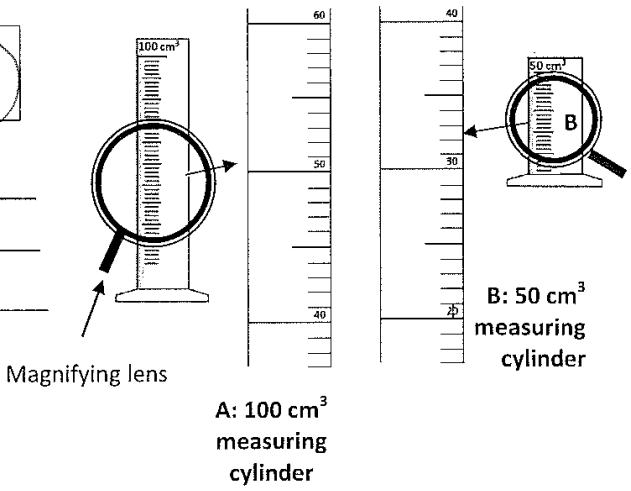
(e) **(i)** Which measuring cylinder would you use to measure 35 cm^3 of salt solution?

Circle your answer

It doesn't matter which	A	B

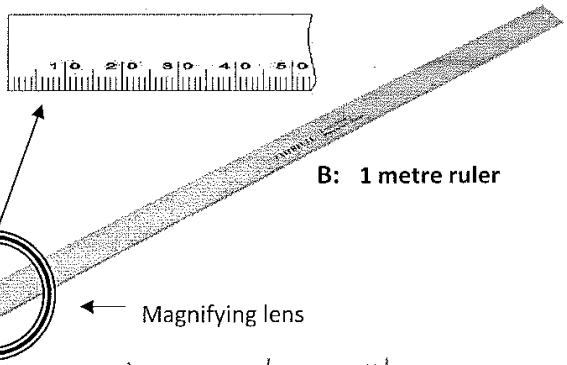
we

(ii) Explain your choice in (e) (i) Can see the measurement accurately



(f) **(i)** Which ruler would be able to measure an 8.0 cm length of string accurately?

A: 10 cm ruler

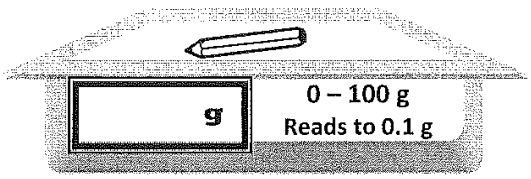


Circle your answer

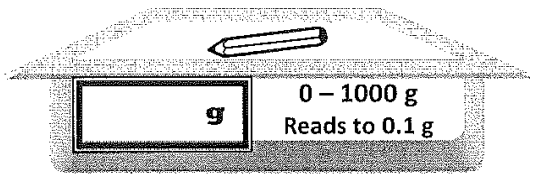
It doesn't matter which	A	B

(ii) Explain your choice in (f) (i) You don't need a ruler with a high value to measure something that is of lesser value.

(g) **(i)** Which digital weighing balance would be able to measure the weight of a pencil more accurately?



A



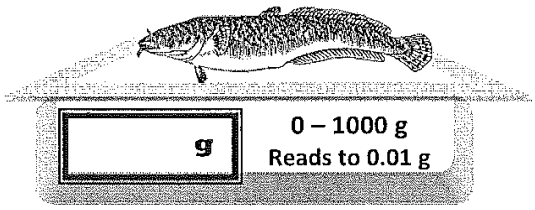
B

Circle your answer

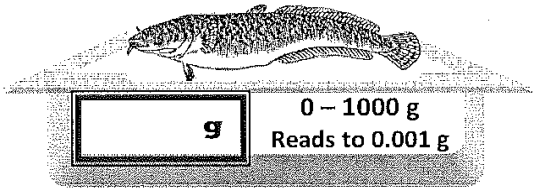
It doesn't matter which	A	B

(ii) Explain your choice in (g) (i) A pencil won't weigh so much till 1000g. A 100g weighing balance would do

(h) **(i)** Which digital weighing balance would be able to measure the weight of the fish best?



A



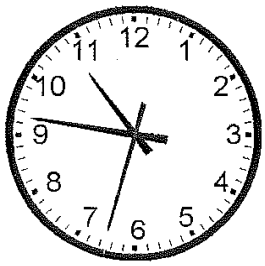
B

Circle your answer

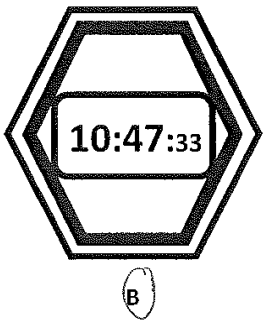
It doesn't matter which	A	B

(ii) Explain your choice in (h) (i) You don't need to measure the weight of the fish accurately

(i) **(i)** Which clock would be able to tell you the time best?



A



B

Circle your answer

It doesn't matter which	A	B

(ii) Explain your choice in (i) (i) We don't need to 'think' about the minute/hour hand to tell the time. The digital watch already

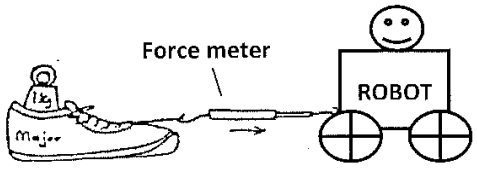
2 Repeats

In an investigation, we often take repeated readings for each value of the independent variable. The statements in the table below are about repeated readings. Think carefully and give your ***best*** response i.e. whether you agree or disagree with the statement.

	Statement	AGREE	DISAGREE
(a)	Three repeated readings are all we need.	A	D
(b)	Most readings when done several times will vary a bit no matter how careful you are.	A	D
(c)	We decide the number of repeats after we have done a few readings.	A	D
(d)	If you get one reading that is very different from all others you should leave it out of your calculations.	A	D
(e)	People who are good at doing experiments always get the same reading each time when making a measurement.	A	D
(f)	The variations in repeated readings are due to human errors only.	A	D

3 The Sole Test

Here are the results of an investigation into how different surfaces affect the 'slippiness' of a trainer. The experiment is carried out by putting a kilogram mass in the shoe and finding the amount of pull needed to drag the shoe along. A remote controlled robot was used for this purpose. Each surface was tested several times and the results for 3 trials are shown below.



Type of Surface	*Pull force (Newtons)*			
	1	*2*	*3*	
Soil on the school's playground	10	10	16	36
Grass on the school field	10	12	14	36
Carpet in the school's library	4	6	6	16
Wooden floor in the school hall	7	6	3	16

(a) Why do you think each surface was tested ***more than once***? Circle your answer.

- **A** to check the first reading
- **B** to see how much the readings vary
- **C** to get the same reading a few times
- **D** to practice and get better

For each of the following questions, choose your answer by circling one of the options, and then explain why you chose that one in the corresponding box.

(b) Looking at the table of readings, answer the following questions.

(i) Which surface needed the ***most*** force to pull the shoe along?

- **A** the grass
- **B** the playground
- **C** I cannot tell which

(b)(ii) Why? ~~Both~~ Some of the experiments have equal results

(iii) Which surface needed the ***least*** force to pull the shoe along?

A the carpet

B the wooden floor

C I cannot tell which

(b)(iv) Why? 2 experiments had the same result unfair. Should repeat experiment.

(c) A similar investigation on tiled floor surfaces found the results in the table below.

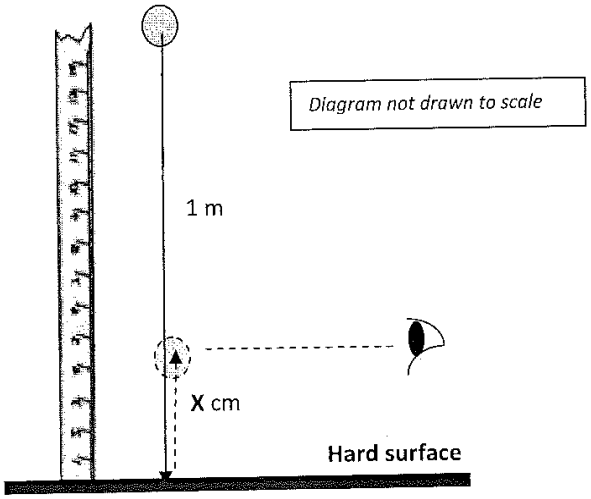
Location of the tiled floor surface:	*Pull force (Newtons)*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
(I) Classroom	14	11	12	13	10	60	
(II) HOD's Office	10	11	13	13	13	10	
(III) Teachers' Common Room	10	11	12	13	14	60	
(IV) Science Laboratory	11	16	11	12	10	60	
(V) School Canteen	10	14	10	14	12	14 / 60	10

For each of the following comparisons, choose the surface that needed ***more*** force to pull the shoe along? Circle your answer and explain why.

No.	Comparison	Circle your answer			Why you choose that one?
(i)	Between (I) Classroom and (II) HOD's Office	(I)	(II)	I can't tell which one	Same result 60N
(ii)	Between (I) Classroom and (III) Teachers' Common Room	(I)	(III)	I can't tell which one	Same result = 60N
(iii)	Between (I) Classroom and (IV) Science Laboratory	(I)	(IV)	I can't tell which one	Same result - 60N. Repeat experiment?
(iv)	Between (I) Classroom and (V) School Canteen	(I)	(V)	I can't tell which one	
(v)	Between (II) HOD's office and (III) Teachers' Common Room	(II)	(III)	I can't tell which one	
(vi)	Between (II) HOD's office and (IV) Science Laboratory	(II)	(IV)	I can't tell which one	
(vii)	Between (II) HOD's office and (V) School Canteen	(II)	(V)	I can't tell which one	all experiments had the same result experiment should be conducted again to ensure fairness.
(viii)	Between (III) Teachers' Common Room and (IV) Science Laboratory	(III)	(IV)	I can't tell which one	
(ix)	Between (III) Teachers' Common Room and (V) School Canteen	(III)	(V)	I can't tell which one	
(x)	Between (IV) Science Laboratory and (V) School Canteen	(IV)	(V)	I can't tell which one	

4 The Bouncing Rubber Ball Test

Jill conducted an experiment to find the bounciest ball; she compared 3 pairs of balls and took 20 readings for all of them. Each time, she would release the ball from a height of one metre and then measured its rebound height in centimetres against a metre rule (see figure below). For easy reference, the balls are referred to as **A**, **B**, **C**, **D**, **E** and **F**.



Study the tables below. If you had done the experiment, tell us ***how many repeats*** you would do to find out which ball was **bouncier**? Circle your answer and explain why you choose that one.

(a) Comparing Ball A and Ball B

Bounce number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ball A	29	29	27	28	29	28	30	26	28	30	25	28	26	27	31	31	30	29	30	28
Ball B	28	30	29	28	29	31	29	27	28	27	28	25	31	26	29	30	29	25	28	27

1 about 3
2 about 10
3 about 20

Because It would be easier to find the average height that the ball ~~bounced~~ bounced

(b) Comparing Ball C and Ball D

Bounce number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ball C	20	21	23	20	21	19	20	23	21	20	19	21	19	17	20	19	22	20	20	21
Ball D	44	43	42	39	40	41	39	40	43	43	39	42	39	40	39	43	38	39	37	43

1 about 3
2 about 10
3 about 20

Because we don't need so many repeats of the experiment. 10 is enough to draw a ~~hasty~~ conclusion.

(c) Comparing Ball E and Ball F

Bounce number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ball E	22	25	27	25	20	14	25	18	20	16	22	20	16	14	22	15	18	11	19	21
Ball F	26	30	26	31	35	28	25	37	32	36	30	36	34	31	35	36	29	35	38	40

1 about 3
2 about 10
3 about 20

Because Too many repeated experiments will cause more human error.

5 (a) Starting an Investigation

(i) You wish to see how independent variable **X**, which has continuous values, affects dependent variable **Y** in an investigation. Which plan shows the sequence you would take your first 4 readings? Circle your answer and explain why?

A

Independent variable **X**	Dependent variable **Y**		
	(1)	(4)	
	(2)		
	(3)		

B

Independent variable **X**	Dependent variable **Y**		
	(1)	(2)	(3)
	(4)		

C

Independent variable **X**	Dependent variable **Y**		
	(1)	(2)	(3)
	(4)		

(ii) Explain why you chose **A, B** or **C**?

We should do the experiment in this order (1st time, 2nd time, 3rd time), then go to the next variable and do the same

5 (b) What next in an Investigation

Each of the tables below shows the first **9 values** you have recorded for a dependent variable **Y** responding to changes of an independent variable **X** in separate experiments. Answer the corresponding question. For the multiple-choice questions, circle your answer and then explain why?

(i)

independent variable X	Dependent variable Y				
	1	2	3	4	5
	F				
20	92	72	84	A	B
30	86	82	80	C	
40	68	70	80		
	D				
	E				

(I) What would be your next **2** readings?

(1) A & B (2) A & C (3) D & E (4) E & F

(II) Why? We should do the experiment such that we use the variable X and repeat the experiment for the 4th time.

(ii)

independent variable X	Dependent variable Y				
	1	2	3	4	5
	F				
20	90	81	77	A	B
30	73	80	80	C	
40	72	79	81		
	D				
	E				

(I) What would be your next **2** readings?

(2) A & B (2) A & C (3) D & E (4) E & F

(II) Why? use variable X (20) and conduct the experiment consecutively.

(iii)

independent variable X	Dependent variable Y				
	1	2	3	4	5
	F				
20	80	70	58	A	B
30	80	78	75	C	
40	80	86	88		
	D				
	E				

(I) What would be your next **2** readings?

(3) A & B (2) A & C (3) D & E (4) E & F

(II) Why? Conduct the experiment consecutively using 1 variable first (20), followed by the next (30)

(iv)

independent variable X	Dependent variable Y				
	1	2	3	4	5
10	A	B	C	D	E
20	***80***	***81***	***82***	F	G
30	***82***	***89***	***83***	H	I
40	***83***	***84***	***85***	J	K
50	L	M	N	O	P
60	Q	R	S	T	U

(I) What do the data (in numbers) show?

The result.

(II) Which letters can represent your next **2** readings and why?

F & G because we do the experiment using the same variable consecutively

(v)

independent variable X	Dependent variable Y				
	1	2	3	4	5
10	A	B	C	D	E
20	***82***	***82***	***83***	F	G
30	***84***	***82***	***81***	H	I
40	***84***	***83***	***80***	J	K
50	L	M	N	O	P
60	Q	R	S	T	U

(I) What do the data (in numbers) show?

The result.

(II) Which letters can represent your next **2** readings and why?

F & G because we do the experiment using the same variable consecutively

End of Questionnaire

Thank you

Annex 7.1 Data for "The Bouncing Ball Test"

Data Set	Ball (categoric IV)	Properties of rubber balls	Mean[1] (of DV repeats)	SD[2]	SE	How many bounces required?
(a)	A	Small differences in mean height in relation to small variations in their rebound heights	28.45	1.67	0.37	about 20
	B		28.20	1.70	0.38	
(b)	C	Large difference in mean heights in relation to small variations in their rebound heights	20.30	1.42	0.32	about 3
	D		40.65	2.06	0.46	
(c)	E	Large difference in mean heights in relation to large variations in their rebound heights	19.5	4.30	0.96	about 10
	F		32.5	4.32	0.97	

[1]The mean value was calculated based on 20 readings.

[2]The standard deviation indicates the size of the variation; the bigger the value, the greater is the degree of variation in the repeated readings.

Annex 8.1: Outline of a Teacher Development Programme for teaching ideas of evidence related to Uncertainty in Measurements

Sequence	Activity	Aim(s)
1	• To complete Questionnaire 2	• Pre-test
2	• Theory: explaining the concepts of evidence • Workshop: Carry out 'spring board man' investigation: how does the mass applied to a 'spring board' affect the height that a 'man' jumps? (See Gott et al., 2008).	• To relate procedural ideas to the overarching concepts of validity and reliability • To develop idea of having to work iteratively in respond to the data being collected (that have a certain amount of uncertainty) - the final set being determined by the level of confidence. During the process of investigation, PSTs will realise the importance of CofEv (e.g. CVs, preliminary trials, etc.) in decision-making
3	<u>Measuring Instruments</u> • Workshop: a range of measuring instruments will be presented to the PSTs. • Theory: Introduction to terms like accuracy, precision, errors, etc. in relation to measuring instruments.	• To understand factors that could affect the accuracy and precision of measurement instruments • To relate accuracy and precision to validity and reliability of measurements
4	<u>Repeated Readings</u> • Workshop: small-group practical involving a categoric IV (for e.g., to measure the absorbency of a paper towel) • Theory: Discussions about repeated readings; the causes of variation; and how variation can be minimised.	• To develop understanding of the intrinsic nature of measurements in that there are always errors, and therefore, uncertainties • To relate to the purpose of repeated readings, the causes of variation, the control of variables
5	<u>Repeated Readings</u> • Workshop: Use of IT activities to develop understanding of descriptive statistics (for e.g.to determine the mean bounce height of a rubber ball; to compare bounce heights of rubber balls) • Theory: Discussions about anomalies, normal distribution, standard deviation and standard error.	• To make meaning of: "the more the readings, the better will be the quantity that is being represented" • To relate to the number of repeated readings, which is being determined by the size of the mean in relation to the degree of variation of the readings
6	<u>Repeated Readings</u> • Workshop: small-group practical involving a continuous IV (for e.g. to plan and find out the relationship between mass and the extension of a steel spring) • Theory: Discussions centre on controlling variables, preliminary trials, intervals, and appropriate range.	• To develop a "repeats-cum-trend-focused" approach towards planning investigations
7	• Workshop: PSTs will be asked to repeat the "spring board man activity", followed by discussions on how they have used the concepts they learned in the programme during the investigation • To complete Questionnaire 2	• Post-test