Master's Thesis:

Measurement and Uncertainty in the First-Year Physics Laboratory; Towards Probing Students' Conceptual Understanding of the Mean

Nuraan Majiet

PHY5000W, Department of Physics, University of Cape Town

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

September 2020

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Declaration

I know the meaning of plagiarism and declare that all of the work in the thesis save for that which is properly acknowledged, is my own.

Nuraan Majiet
30th September 2020

Abstract

Physics is about sense-making. The world we live in and experience through our sensory modalities is highly complex. In order to make sense of this complexity we reduce the experiences to a more simplified form. The way in which this is achieved is through modelling. Physics consists of both theory and experiment, thus modelling in physics consists of two components: (1) conceptualization and mathematization (theory) which involves ontological innovation and introducing variables and (2) designing experiments which leads to measurements (experiment). We can then compare our theoretical predictions with our measurements. The present work is primarily focused on aspect (2) dealing with the modelling of experiment.

First year physics courses include a teaching component directed at the key aspects that relate to experimentation. This includes the key concepts with regard to measurement and uncertainty. However, these have proved to be challenging aspects of a first-year curriculum and students often resort to rote methods. Student understanding of measurement and uncertainty was explored in detail in a series of studies that were carried by a collaboration between UCT and the University of York. This work showed that students exhibited a wide variety of ideas with regard to all aspects regarding data, ranging from data collection to data processing. Based on their theoretical constructs that explained this variation in terms of point and set paradigms, they concluded that the purpose of teaching was to move students from the point to the set paradigm. Despite the fact that they created an instrument (Physics Measurement Questionnaire (PMQ)) to measure such a shift, it is not clear that the measured shift reflects actual conceptual change. This is particularly so insofar as combining multiple readings into a single number such as the mean is concerned. While many of the questions on the PMQ do attempt to probe student thinking, the question regarding the mean is in fact purely calculational. Therefore, the nature of the responses does not allow one to fully determine to what extent the calculation follows from an appropriate model or whether it is simply an arithmetic step that is carried out *without any model in mind.* While calculating the mean might be regarded as a step forward for students who were previously classified as point thinkers it can be argued that this is in fact a retrograde step from a modelling perspective in that the step can be described as “model abandonment”. Thus, rather than the mean being a stepping stone to further understanding of uncertainty, it could in fact prevent such a learning trajectory. As seen from the PMQ it is not easy to pose questions that probe what model, *if any,* students have in mind when calculating the mean.

The present work thus aimed to explore the degree to which it was possible to identify to what extent students used the mean with some model in mind. The starting point for the work was the PMQ. Questions were posed in the same manner but with the aim of eliciting the reasons why students perceive the mean to be the appropriate way to proceed during data analysis. To what extent is it possible to probe students’ modelling approaches in the first-year laboratory? Is it possible to design a non- interview methodology in order to identify their reasons for using the mean? To investigate this a number of questions were constructed and administered to two small groups of students, 20 and 30 respectively, as part of a two step iterative developmental research process.

The questions were administered to first-year physics students at the University of Cape Town. The final questionnaire consisted of four questions. The two data collection probes were taken directly from the PMQ and placed at the beginning of the questionnaire for control purposes and the two pilot questions were adapted from the Using Repeated Distance (UR) Probe in the PMQ. UR was reformulated into two questions with an explanation component; one question investigated what students use as the final result in a purported experiment and the other looked at what they predicted as the next value.

The analysis comprised careful investigation as to the "Level of Informativeness" provided by the questions followed by a cross probe analysis where the Level of Informativeness allowed for this to be done.

The present studies that were carried out indicated that there was no straightforward way to elicit information as to whether the student had some model in mind or not. However, a number of insights into the way forward were gained. These included the way in which questions could be framed around the issues of the mean that allowed for some level of inference to be made. While some further work still remains insofar as this is concerned, we suggest that these questions be included in future versions of the PMQ.

Acknowledgements

I would like to thank my parents, Bahidja and Moekriem Majiet, as well as my siblings for their continued love and support throughout my master's degree studies. It is through their constant encouragement, interest and enthusiasm that I have been able to overcome obstacles and celebrate my successes.

I would like to thank my Supervisor, Saalih Allie, for believing in me and providing me with the opportunity to pursue my master's degree, in particular, in the field of Physics Education Research. Our often reflective and contemplative discussions (and arguments) have been invaluable to my experiences and growth these past few years. I am also grateful for the other opportunities I have been afforded through Saalih's support. I have no doubt that my travelling and teaching experience will be instrumental in my future career pursuits.

I would like to thank the Center for Higher Education and Development(CHED) and the Department of Physics (and HOD Andy Buffler) for financial support with my master's degree as well as funding my travels to AAPT, PERC and FFPER.

Thanks to the Department of Physics' students and staff for their participation in my research.

I have had the privilege of meeting and working with people from all walks of life. The Physics and Astronomy Education Research Group (PhAsER) has been a constant support system and the source of many fruitful discussions. Thanks to Dale Taylor, Tshiamiso Makwela, Mayhew Steyn, Isiaku Mbela and Alex Sivitilli. Warm thanks to all the friends and colleagues I've met through AAPT, PERC and FFPER. I am grateful for all the conversations, feedback, comments and suggestions I have received over the years. Thanks to Mary Grace McGeehan for her pleasant company and meaningful discussions during travels as well as for her editing and proofreading of this work.

I am also grateful to the Department of physics staff and postgrad students for allowing my work environment to be a pleasant one.

Table of Contents

List of figures

List of tables

Chapter 1: Introduction

1.1 Making sense of the complex real world

The world we live in and experience through our sensory modalities is highly complex. In order to make sense of this complexity we reduce our experiences to more simplified situations that allow us to identify causes and essences more easily, and thereby lead to greater understanding. Fig 1.1 shows this overall approach of "modelling" in schematic form.

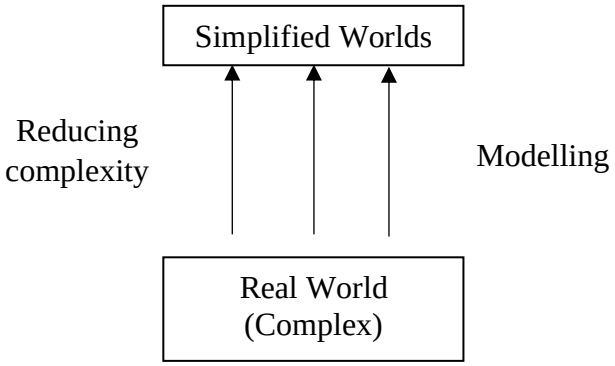


Figure 1.1: General approach to reducing the complex real world (experiences) to a number of simplified worlds that describe the explanatory aspects through a process called modelling.

Physics offers a particular approach to modelling that is based on two inter-linked components: (a) theory and (b) experiment.

(a) The theoretical component proceeds as follows:
Complex situation → reduced complexity → concepts creation[1] → identification of a mathematical variable within a mathematised description that finally allows for numerical predictions

(b) The experimental component proceeds as follows:
Complex situation → reduced complexity → "experiment" → measurements

The level of agreement between what is predicted by theory and what is measured by experiment forms the basis of the physics approach. Fig 1.2 shows this dynamic relationship between the two "arms" of physics. This approach regarding the enterprise of physics has been described in detail by Hestenes (1992) and is discussed in more detail below.

[1] Di Sessa [DBER paper] refers to the creation of idealized conceptual constructs that form the basis of the theoretical physics world as "ontological innovation".

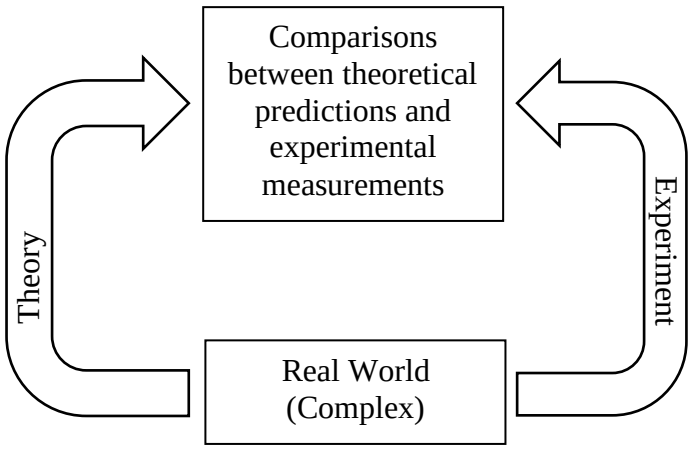


Figure 1.2: The two components of physics modelling: (a) theory and (b) experiment. The overall aim is to refine both theory and experiment so that the resulting description is consistent with that of experiment.

1.2 Physics Modelling: *Theory*

In his 1992 paper Hestenes (Hestenes, 1992) emphasizes the clear distinction between the "Physical World" and the "Conceptual World". These two distinct domains are linked through the modelling process in which the complex Physical World is reduced to theoretical conceptual elements and inter-relationships in the "Conceptual World". The entities in the "Conceptual World" are cognitive creations and while they may be derived from physical entities, they are idealizations with their own clearly defined properties e.g. a point mass or an ideal resistor. By way of example the paper uses Newtonian Physics (see Fig. 1.3) to illustrate the key points of the modelling paradigm.

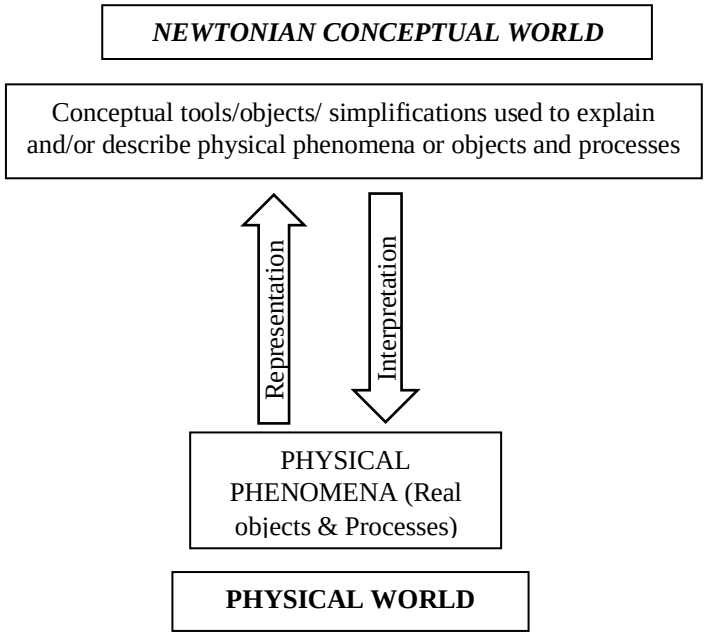


Figure 1.3: Adapted from Hestenes (1992) (Fig 1). This figure shows that the Newtonian World is a Conceptual World of reduced complexity which can be used for modelling actual objects and processes in the Physical World.

Thus, the "Physical World" consists of physical phenomena (real objects and processes) that are linked to the Newtonian World that consists of models of physical phenomena as shown in Fig1.3 (reproduced from Fig.1 in Hestenes (1992)). The detailed example that is provided can be thought of as a specific case of the branch of physics known as "theoretical physics".

1.3 Physics Modelling: *Experiment*

The previous arguments that apply to theoretical physics apply equally to experimental physics. As pointed in Hestenes (1992), experiment consists of two aspects, firstly, (a) the *procedures* that are carried out that lead to data production and secondly, (b) the analysis and interpretation of these data. With regard to the latter Hestenes notes that "*without analysis and interpretation, the data* [collected by the experimental procedures] *are meaningless*".

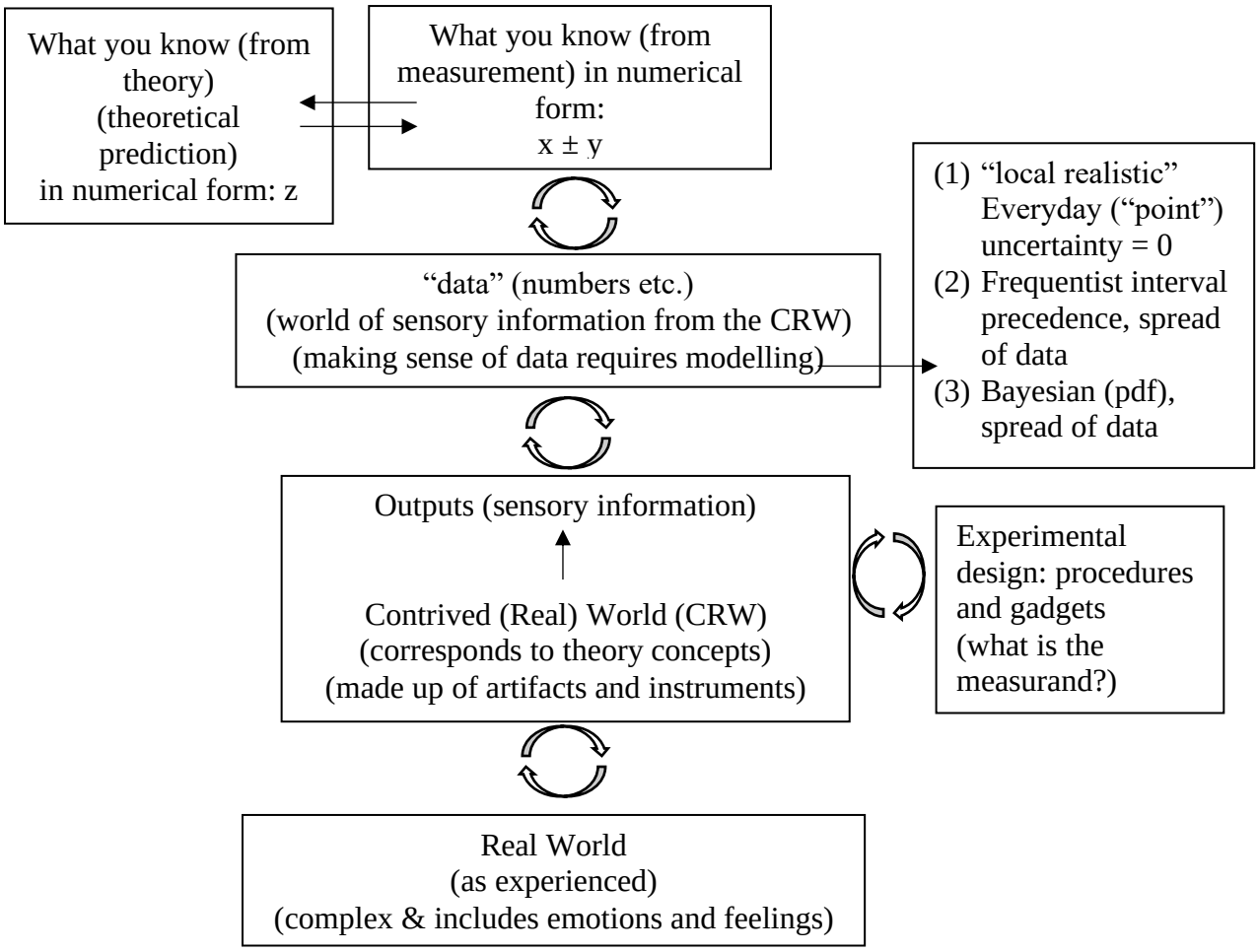


Figure 1.4: The physics modelling pathway of experiment can be considered a two-step modelling process: (1) which results in a contrived (real) world (experimental design, procedures and gadgets) and (2) which deals with measurement (data analysis and interpretation). Step two is involved in making sense of observations so that the complex data can be put in a form that allows comparisons to be made with the theoretical prediction.

1.4 Issues pertaining to teaching physics from a modelling perspective

The idea that physics is about modelling is well understood by physicists but is seldom foregrounded in teaching. Thus, it is not clear that students understand this distinction and more often than not appear to conflate the "Physical World" and the "Conceptual World". For example, if we look at typical first year physics textbooks (Knight, Halliday and Resnick, Giancoli etc.) and single out the kinematics chapter, examples of problems and solutions tend to conflate what happens in the "Physical World" with the "Conceptual World". For example, cars and buses, which are extended bodies in the "Physical World", are conflated with point particles which are objects that only exist in the "Conceptual World" (and whose behavior can be described mathematically).

A similar situation can be found in the way in which the experimental aspects of physics are taught. For example, introductory physics laboratory work is often reduced to a cookbook approach in which the purpose of the exercise is to prove some aspect of theory. An important part of experimental physics deals with measurement and uncertainty. However, while tasks carried out in the laboratory involve modelling and theoretical constructs, this is not explicitly known by students. More pointedly for the present work, students process data without knowing that they are using a particular modelling approach to make decisions and perform data reduction calculations. Once data have been collected, we need a modelling approach to be able to make decisions on how to proceed. Thus, we often teach students that they should take several readings in an experimental situation and then make them calculate a mean. However, this does not make it clear that a particular modelling approach is being used (the frequentist) and often resorts to rules of thumb to deal with a single reading as pointed out in Allie *et al.* (1998).

1.5 Research into student understanding of measurement and uncertainty

The work carried out at UCT focused on experimental work; in particular the analysis and interpretation of data. Over the past few decades a number of studies have been carried out regarding students understanding of measurement and uncertainty. This section will be divided into four parts; (1) research done by other groups, (2) work carried out prior to the UCT and University of York (UCT-York) collaboration that is related to the UCT-York research, (3) the work carried out by the UCT-York collaboration and (4) the focus of the present study.

1.5.1 Research into student understanding of measurement and uncertainty by other groups

Sere *et al.* (1993) investigated first-year university students' conceptions about measurement. The students completed a theoretical (statistics) course, and 8 hours (divided into two sessions) of laboratory work in the context of optics and electricity. The students were then required to take a written test at the end of the course. The authors found that even after instruction students had a poor understanding of the procedures and the advantages of statistics. Their findings included that students did not have a good grasp on why certain experiments require several measurements to be made. This lack of understanding meant that the actions and reasoning students took when they repeated measurements were not aligned with the theory. Some action and reasoning observed included giving data points a hierarchy dependent on the order of

collection of data points. Students often only repeated measurements to validate the first data point.

Zanari and Miller (2003) looked at how students aged 9, 11, and 13 reason as they collect and evaluate experimental data and draw conclusions about the relationships between variables. They designed two experiments; (1) an experiment where an independent variable co-varies with the dependent variable and (2) an experiment where theoretically the independent variable does not co-vary with the dependent variable. They found that the students were less likely to repeat measurement for experiment (1) compared with experiment (2). The students only repeated measurements when they got an unexpected result, or to check or confirm their initial measurements. Most students replaced their measurement with a repeated measurement and did not record the individual measurements. The majority of students in the study did not refer to an average of repeated measurements but drew conclusions directly from the individual measurements and argued that the variation in the data was due to error.

Masnick and Morris (2002) conducted individual interviews with primary school (ages 9 to 12) and undergraduate students (average age 20). They looked at the students' confidence when comparing data sets that varied in sample size (one to six data points), consistency in overlapping data pairs (from zero to two) and variability relative to the mean. The students were presented with tables of data related to the performance of two athletes. The students were asked what conclusions they could draw from the information, the reasons for these conclusions and how certain they were. The students were also asked to predict the next data point for each athlete, and how certain they were about the difference between two predicted values. The study's findings varied depending on the students' schooling level. They found that undergraduate students were significantly more confident when there were more data points and the primary school students were more confident with smaller sample sizes. They also found that primary school students drew conclusions after only seeing one pair of data while the undergraduate students become more confident about their conclusions after seeing at least four consistent pairs of data.

Masnick and Klahr (2003) investigated primary school students' understanding of experimental error. They found that while primary school students were able to recognize potential sources of error in an experiment, they were unable to completely integrate what they know about experimental error coherently. They also found that for students of different age groups, there was a gradual developmental shift as the students learned more about experimentation and causation.

In a French study of high school students (14-17 years), Coelho and Séré (1998) described students' tendencies and difficulties when carrying out various activities involving measurement. The authors conducted interviews to gain insight into the students understanding of the following areas: (1) data collection, (2) data processing and interpretation, (3) the meaning given to measurement and (4) the capacity to draw conclusions from data. In the study, the authors described the students' search for the 'true value' of a quantity, and their dissatisfaction with the inconsistency of their measurements.

Evangelinos *et al.* (1998) found that undergraduate students had deeply rooted views about 'exactness' and 'precision' that acted as barriers to their acceptance of uncertainty as an intrinsic property of scientific measurements. Even after instruction, many students would retain the view that a single measurement taken with a laboratory instrument could give the true value of a measurand (the quantity being measured). Séré (1993) and her colleagues also observed that students were very loose in their use of terms such as 'precision', 'accuracy' and 'systematic' and 'random errors'. Evangelinos *et al.* (2002) reported on an intervention study with first year university students in Greece. They categorized their students as being 'exact', 'approximate' or 'interval' reasoners with regard to their views on the relationship between theory and data. Ryder and Leach (1999) and Leach *et al.* (2000) also investigated students' view on the relationship between theory and data.

Vellom and Anderson (1999) looked at how grade 6 students validate their experimental findings to their peers. The students had to work in groups and therefore needed to use strategies for coming to a consensus. The study showed that social and academic status played a role in their decisions. A focus on the nature of the experimentation was also shown to be important. This included discussions about experimental techniques and the need to replicate data. Fairbrother and Hackling (1997) discussed the closed nature of traditional laboratory courses and how this stems from the epistemological view of science as a collection of facts. They argue that closed laboratory tasks reinforce students' expectations of the existence of a 'right answer' to any experimental problem. This means that when students get inconsistent or unexpected measurements, they think they have made an error rather than consider uncertainties and what this means with regard to the measurement process.

Some later studies explored students' ideas about expectation and variation. Leavy and O'Loughlin (2006) conducted a study in Ireland that investigated preservice elementary teachers' understanding of the mean. Data was collected through a written instrument as well as through clinical interviews. 25 out of 263 participants were interviewed based on their written responses for the instrument. The instrument consisted of five mathematical tasks. Task 1 required the comparison of two unequal sized data sets, task 2 required computation of a weighted mean, tasks 3 and 4 looked at the relationship between the mean and the data sets from which the mean is constructed, while task 5 investigated the possession of visual or kinesthetic understanding of the mean. The study showed that while the students demonstrated mastery of computational skills related to the mean, the students had difficulty applying their knowledge of the mean to unfamiliar tasks. Watson (2018) explored the primitive understanding of expectation (e.g. mean) and variation (e.g. standard deviation) of six-year-olds as they worked through four hands-on tasks that were devised for older students. The children were also asked to make predictions within the context of these four tasks. This study was based on previous research in which Watson (2005) claimed that children develop the concept of variation before that of expectation. Watson (2018) found that variation either created the student's prediction or provided supporting evidence that the student's expectation was reasonable and claimed that appreciation of variation is the starting point for children's engagement with the practice of statistics.

Other studies looked at the philosophy of science and how students' perceptions about science and experimentation can affect their engagement with the laboratory. Hammer (1994) investigated the epistemological beliefs of a small group of undergraduate physics students in

three areas and concluded that their beliefs affected their success in physics. In particular, he studied their (1) beliefs about the structure of physics as (a) a collection of isolated pieces or (b) a coherent system, (2) beliefs about the content of physics as (a) formulas or (b) concepts that underlie the formulas and (3) beliefs about the process of learning physics as (a) receiving information or (b) an active process of reconstructing one's understanding. Tsai (1997) investigated the scientific epistemological beliefs and learning orientations in a group of Taiwanese eighth graders. Wilcox and Lewandowski (2018) investigated university students' beliefs about the nature of experiment, while Ibrahim (2009) and Buffler *et al*. (2009) looked at the relationship between students' views of the nature of science and how it affected their views of the nature of scientific measurement. Tlowana (2016) looked at how emotions and cognition affect students' perceptions and engagement with laboratory tasks. Lippmann (2004) looked at students' argumentative and decision-making skills for data gathering and analysis, in particular student frames (mind-sets) which are necessary for productive use of these skills.

1.5.2 Prior research related to the UCT-York studies

The following account is based on references: [18-20, 24, 32 & 33].

In 1994 Gott, Duggan, Lubben and Miller probed student understanding of measurement in the UK. This work was done under the Procedural and Conceptual Knowledge in Science (PACKS) Project. The main aim of the work was to develop a model that linked students' performance of investigative tasks to their understanding of measurement. In order to investigate this, they constructed a written instrument that was administered to middle school students (aged 11-15). The key themes that were probed by the instrument were *data collection, data processing and data comparison.* The key features of the instrument are that (a) short scenarios regarding different aspects of measurement are described in words, (b) each question is based on a different scientific context – physics, biology etc. and (c) a number of different opinions are voiced by a different characters (sketches). Respondents were asked to select an opinion that was most closely aligned with their own opinion and then to provide a written explanation of their choice. An example of a question taken from the instrument is shown in figure 1.5.

The main result from the work was Lubben and Millar's (1996) '*Model of progression of ideas concerning experimental data'* which identified a hierarchy of student ideas about measurement (see Figure 1.6). The categories A-H that were proposed to describe student progression of understanding of the collection and evaluation of empirical data is discussed here. In the model, A is the least sophisticated category and H the most. In category A students regarded that only one careful measurement had to be taken. The students in category B had the view that repeating measurements could only lead to further issues when one gets a different value. Therefore, only one measurement is necessary. Students in category C saw the sources of scatter as only due to inadequate equipment usage. Practicing first will ensure one measures the correct value. In category D, similar to C, students regarded the scatter as only due to inadequate equipment usage, however for this category, the students think repeating measurements will give one the same result each time. This validates that careful measurements have been made. In category E: the students think that one has to take the average but this would give the same value continuously, so the experiment has to be manipulated to create a scatter. For category F the mean is calculated to account for the variation which is seen as due to inaccurate measuring. For

this category only an authority figure can judge the quality of the result. The students in category G obtain a mean and use the spread of the measurements as an indication of the quality of the result. In category H the students are able to judge the consistency of the set of measurements and reject anomalous measurements before calculating a mean.

The authors emphasized that progression through the levels does not reflect the students' progressive learning paths. However, the model provided a tool for classifying students' measurement actions in terms of the underlying measurement ideas.

QUESTION 1

This question is about an investigation to find out how quickly sugar dissolves in water. Pupils in a class make these measurements:

- time how long it takes for sugar to dissolve
- using 5 g of ordinary white sugar
- and 100 ml of water
- at 80˚C

Jim, peter and Mark worked as a group. They measured the time for the sugar to dissolve. They find it takes 32 seconds.

Who do you agree with?
Explain why you agree with him.

Figure 1.5: A question taken from the paper "Ideas About the Reliability of Experimental Data" by Lubben and Miller (1995). This showcases the format used for the questions in this prior work.

Table 5. Levels of students' understanding of the collection and evaluation of empirical data.

Level	*View of the process of measuring*	*How to evaluate your result*	*What to do with readings which differ appreciably from most of the others*
A	Measurement is straightforward; measure once and you will get the right value.	Not an issue. A measurement is correct.	Not an issue.
B	Measure once and take this as the right value. Repeating will lead to a different result, but any result is likely to be as good as any other, so there is no point in repeating. In fact, it only confuses the issue.	Unless something has obviously gone wrong, a measurement is correct. In familiar contexts, it should be close to what you would expect for the quantity measured.	Not an issue.
C	If you have adequate equipment and use it carefully, your measurement will be right. Make a few trial measurements to get familiar with the equipment; then take the measurement you want.	Unless something has obviously gone wrong, a measurement taken after a few trial runs will be correct. In familiar contexts, it should be close to what you would expect for the quantity measured.	Ignore. (If a 'trial' reading is different, this is due to lack of practice; if 'final' reading is different, this is the result of practice.)
D	If you have adequate equipment and use it carefully, your measurement will be right. Repeat measurements in order to get the same result twice.	Getting the same value twice shows you have measured carefully enough.	Ignore.
E	You should repeat a measurement and take the average. But repeating *exactly* the same measurement will lead to the same value; so change the conditions a little each time.	Variation is to be expected. Not an issue.	Variation is to be expected. Include all values in calculating an average.
F	Careful measurements may get close to the right value of the quantity you are measuring but you can never be sure you have found it. Taking an average of several measurements allows for this.	Cannot be evaluated from 'inside'. Only method is to check with an authority source (the teacher, or a textbook or databook).	Including them when calculating the average will take care of them. In fact, this is why we use the average.
G	– as above –	Can be evaluated from 'inside'. The spread of the measurements is an indication.	– as above –
H	– as above –	– as above –	It is appropriate to exercise a judgement about the set of data, and reject anomalous results before taking an average. The mean of some sets of data, therefore, may be better than of others.

Figure 1.6: Lubben and Millar's (1996) 'Model of progression of ideas concerning experimental data' which identified a hierarchy of student ideas about measurement.

1.5.3 UCT-York studies

The following account is based on the following references: [1-13, 30, 35, 37, 40 & 41].

Following on from the work described in the previous section, a collaboration was established between the University of Cape Town and the University of York in 1995. The collaborative group aimed to investigate and interpret first-year physics students' understanding of measurement and uncertainty. Part of this included the development of a theoretical basis for the construction and implementation of a new introductory physics laboratory curriculum. The group not only aimed to facilitate the development of students' abilities in performing experimental procedures and using the tools of analysis, but also aimed to deepen students' understanding of the nature of measurement and uncertainty. This subsection describes the key studies carried out by the collaborative group and the culmination of the work in terms of the Point and Set paradigms.

In order to investigate students' understanding of measurement and uncertainty, a process of developing an instrument that was appropriate for older and more educationally advanced students was started. The UCT-York collaboration modelled their instrument on the one described in the previous section. However, they made a number of key changes to the instrument since the group of students at UCT had different backgrounds from the ones in the PACKS study. The changes made to the instrument is described below.

Many of the South African students were English second or third language speakers, therefore the language structure and vocabulary had to be carefully chosen for each of the questions in the instrument. Next, neutral cartoon characters (as seen in figure 1.8) were used instead of sketches of people to avoid bias caused by ethnicity, culture and/or gender. Letters were also used as descriptors rather than names for similar reasons. The authors conducted interviews to confirm that the language used was at an accessible level. The preference for the neutral cartoon characters were also validated through interviews and written feedback. The authors also showed that the majority of students found the cartoons to be gender neutral and race free.

The PACKS study used different contexts for each question. As this could lead to confusion for the South African students, the developed instrument was constructed around a single physics scenario from which all the questions would follow.

An experiment is being performed by students in the Physics Laboratory.

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

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

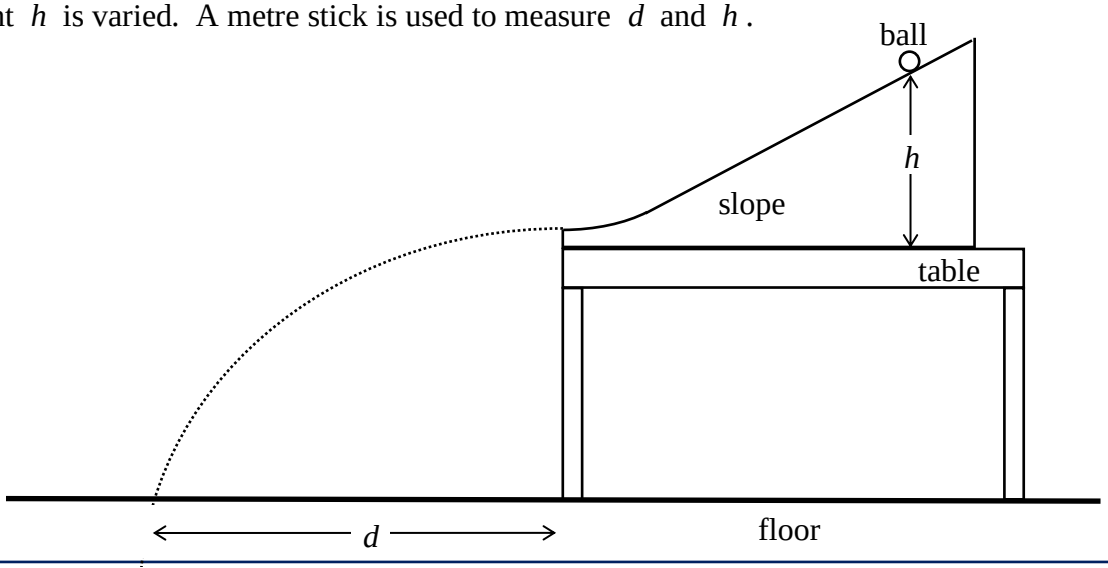


Figure 1.7: The experimental context used for the probes in previous studies. It consisted of a ball rolling down a ramp.

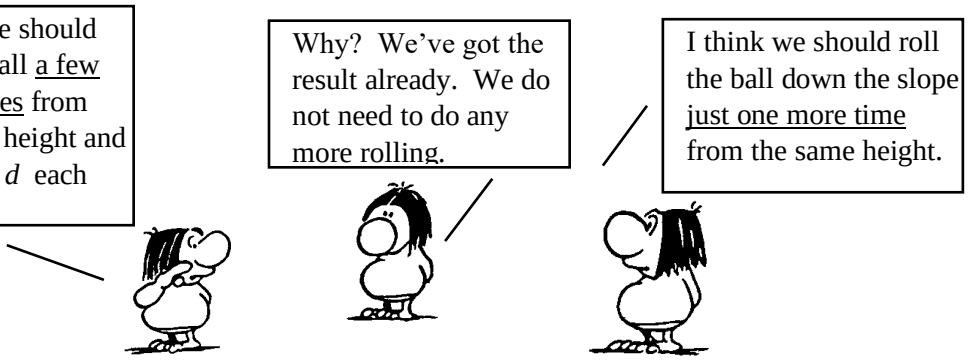
All the probes followed the same style and related to the experimental context as shown in figure 1.7. The probes also followed a deliberate sequence which mimicked the order of decisions made in experimental work. Each probe introduced a practical laboratory situation which required a decision to be made. Similar to the instrument described in the previous section they framed all of the probes in the form of a discussion. Some probes then required the respondents to select the opinion with which they most closely agreed while others were open ended. For the probes with options, the suggested actions were selected such that the respondents could have chosen them for a variety of reasons. For many of the probes, one option also allowed the respondents to have a different opinion.

A more recent version of the instrument is now known as the Physics Measurement Questionnaire (PMQ) and can be found in Appendix 3 as well as on Physport.[2].

Repeating distance probe

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

The following discussion then takes place between the students.



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

Explain your choice

Figure 1.8: An example of the format and structure of the questions used in the UCT-York studies, in particular, the Repeating Distance Probe taken from the Physics Measurement Questionnaire.

The PMQ probed the same key themes as the PACKS study. The PMQ consisted of 8 questions. Three related to *data collection* (Repeating Distance (RD), Repeating Distance Again (RDA) and Repeating Time (RT)). Two questions related to *data processing* (Using Repeats (UR) and Straight Line Graph (SLG)) while two questions related to *data comparison* (Same Mean Different Spread (SMDS) and Different Mean Overlapping Spread (DMOS)). It is interesting to

[22] Physport is a website where physics academics can find resources based on physics education research (PER) that can support teaching (www.physport.org). Sarah "Sam" McKagan (AAPT) is the director of the website.

note that one of the original data processing questions which dealt with an anomalous measurement result was taken out of the final version of the PMQ. The strict ordering of questions was deliberate so as to prevent later probes from offering hints at the "correct" response or creating bias in the students' reasoning. Students were also discouraged from going back and changing their responses. However, at the end of the questionnaire they were allowed to specify what they would change if they could.

The results that followed from administering this instrument to a group of physics first year students were analyzed using a grounded approach and then distributed according to the Lubben and Miller's Model described in the previous section. Only a few students could be classified within the levels A, C and D, and this was not consistent. While many students could be categorized within the levels F, G, or H, those who could be classified consistently demonstrated greater sophistication than was allowed by the Lubben and Miller Model. Thus, owing to the fact that the group of students were more academically advanced (first year university) relative to the UK group (middle school students) it was found to be necessary to add one category to Lubben and Miller's Model. Students in this category (I) understood that a mean together with a measure of spread form an interval that may be used to judge the consistency of data ensembles. Another limitation of Lubben and Miller's Model for the UCT students was the ranking of student responses as sophisticated even when there was haphazard use of terminology in their explanations.

A key theoretical advance in the field took place when the group proposed that all student responses could be attributed to two underlying perspectives regarding data: as either providing a single 'true value' or a 'spread' of values. They termed these the Point and Set Paradigms[3] respectively. Based on a series of papers that used these constructs the key ideas that characterize the point and set paradigms are described below:

"The **point paradigm** is characterized by the notion that each measurement yields either the correct (true) value or an incorrect value of the quantity being measured (the measurand). As a consequence each measurement is regarded as independent of the others, except to confirm or reject a specific value, and individual readings are not combined in any way. This way of thinking also manifests itself in the belief that only a single (very careful) measurement is required to establish the true value. If an *ensemble* of readings with dispersion does emerge, representations of the measurand are based on the individual data points only, such as for example, the selection of a recurring value in the data set or a one-to-one comparison of data values between different data sets."

"The **set paradigm** is characterized by the notion that each reading is an approximation of the measurand and that knowledge about the measurand can never be perfect in principle. The most information regarding the measurand is obtained by using all available data to construct distributions from which the best approximation of the measurand and an interval of uncertainty are derived."

[3] For the context of this work the term paradigm is used as described by Kuhn (1970): a '… constellation of beliefs, values, techniques and so on shared by members of a given community'. Therefore, the two paradigms underlie both student reasoning and actions associated with measurements.

In nearly all practical situations in the introductory laboratory, the best approximation of the measurand will either be the reading itself (in the case of a single reading) or the calculated average value of a set of repeated readings. Thus, the key difference between the two paradigms is that students using the point paradigm draw conclusions about the measurand directly from individual data points, while those using the set paradigm draw conclusions about the measurand from the properties of the distribution constructed from the whole ensemble of available data.

Table 1.1: Table of Point and Set Paradigms as taken from Monograph: Teaching scientific measurement at university: understanding students' ideas and laboratory curriculum reform.

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

In later studies, the authors then classified students in terms of how consistently they used either the point or set paradigms. From this they inferred the students' level of understanding of measurement uncertainty. A student who was firmly located within the set paradigm was regarded as having a good understanding.

After classifying student responses according to the point and set paradigms it was found that the majority of university entering South African students, who are generally underprepared for university, can be characterized as thinking within the point paradigm before instruction. Previous data also showed that only a small percentage of students could be consistently categorized within the set paradigm after a traditional[4] laboratory course. Therefore, the group started a development project aimed at designing a curriculum that would suit the diverse needs of UCT students and facilitate a shift from the point paradigm to the set paradigm. This project led to the development and implementation of the Probabilistic Approach to Measurement. The authors argued that part of the problem with traditional laboratory courses has to do with the theories of data analysis that are used. The modelling approaches that have been used to interpret data have been the subject of much discussion over the past few years. Essentially the debate has been over the Frequentist approach versus the Bayesian. The differences between the two approaches are most pronounced when it comes to how to deal with a single measurement. Thus, the Frequentist modelling approach of how to deal with data does not extend to dealing with a single measurement. In the limit of large numbers of data however, neither modelling approach has any apparent advantage over the other and the mathematical descriptions are in fact the same. However, the interpretations of these descriptions are not the same. The deep difference between

[4] Cookbook style laboratories.

the two modelling approaches can be summarized as follows: “in the Frequentist modelling approach the phenomena are certain and the data are uncertain while in the Bayesian modelling approach the data are certain and the phenomena are uncertain”. This has implications for how students are taught to make sense of the spread in data. The point and set paradigms can be regarded as modelling approaches that are used (often implicitly) by students. Traditional instruction typically uses the Frequentist modelling approach while the Probabilistic approach to measurement includes Bayesian notions as recommended by the International Standards Organization in the published document the “Guide to Expressing Uncertainty in Measurement” (GUM).

The authors claimed that the role of teaching should be viewed as shifting students from the point to the set paradigm. Based on the PMQ, using the Probabilistic approach to measurement they were able to see improvement in student shifts from the point to the set paradigm. However, this is an observational outcome. While the PMQ is informative in terms of measuring how many students are consistent point and set thinkers pre- and post-instruction, it is not designed to establish if this was due to conceptual change or if this is the result of learning set-like actions without a deep understanding of why set-like actions are more appropriate than point-like actions.

1.6 The present work

The present work forms part of the research into student understanding of measurement and uncertainty described in the previous section. In particular, the study attempts to investigate how students think about data obtained from measurement at a more fine-grained level than that obtained from the Physics Measurement Questionnaire. More specifically, do measured student shifts from a point to a set paradigm come about due to actual conceptual change or are they the result of recognizing familiar situations that can successfully be processed according to appropriate prescriptions?

As discussed previously, data handling also involves modelling. From the present perspective, the point and set paradigms are what can be observed and reported upon as shown in the previous section. In general, we view a paradigm as being associated with the model[5]. The work of UCT-York can be regarded as using the paradigms as associated with a particular model. In this case the point paradigm is associated with the intuitive model that is used in measurement in the same way that the Aristotelian model gives rise to the Aristotelian paradigm. On the other hand, the set paradigm is associated with a scientifically accepted model of dealing with data. It is often the case with students that they appear to be following a particular model when classifying them according to their paradigmatic “moves”. However, it may be the case that the move is carried out without actually following from the model but rather learnt by rote.

For example, the use of the mean is associated with the “set” modelling of data – the key being that the data are modelled by a mathematical function which has to be symmetric; the complex point space is modelled by a simple pattern. The properties of the pattern are then used to convey

[5] The notion of paradigms is regarded by Kuhn (1970) as the “tools and stuff” that are used within a model.

the summative outcome of the measurement. This allows for the most representative proxy to be the mean.

In previous work the role of teaching is viewed as shifting students from the point to the set paradigm. The authors argued that after using the Probabilistic approach to measurement they were able to see student shifts from the point to the set paradigm. However, this is an observational outcome and may not be an appropriate measure when inferring which modelling approach students have in mind when carrying out tasks. In particular, it remains unclear that students are indeed using the mean following on from the "set" modelling of data rather than as a calculational tool learnt through rote.

The present study answers the following research questions:

(1) What reasons do respondents give when justifying the calculation of a mean?
(2) To what extent can we create a modified version of the PMQ that better elicits respondents' reasoning when calculating the mean of a set of data?
(3) To what extent does respondents' reasoning about means relate to other evidence for set or point like reasoning?

In order to probe these finer grained aspects of student understanding, a two-step process of developing a suitable questionnaire was initiated. The following section outlines this process.

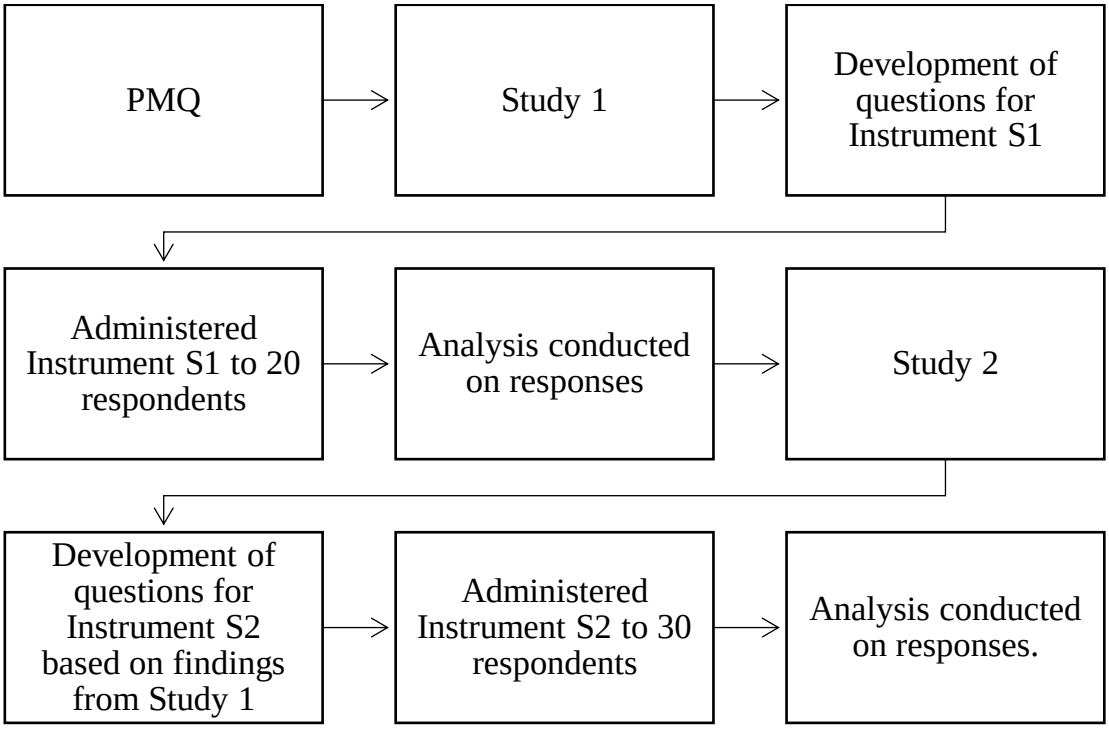
1.6.1 Overview of the present work: scope and focus

The present work consists of two studies that followed each other: Study 1 and Study 2. The PMQ was taken as the starting point for the development of questions which could probe the relevant area of focus as described in the previous section. The PMQ focused on three areas of laboratory work: data collection, data processing and data comparison. However, for the purposes of the present work, the relevant probes were the ones focused on data processing as these related to students' ideas regarding the mean. The relevant PMQ questions were then adapted and developed in the two studies as described in detail in chapters 2 and 3. The PMQ questions that were used in the present work is shown in the table below.

Table 1.2: Overview of the development of questions attempting to probe students' conceptual understanding of the mean

Relevant probes selected from the PMQ and used in the present work	
Repeating Distance (RD)	Used for control purposes
Repeating Distance Again (RDA)	Used for control purposes
Using Repeated Distance Measurements (UR) (Open-ended calculation)	Adapted in the present work
Straight Line Graph (SLG) (Open-ended with diagram)	Adapted in the present work

While case studies, which is a well-known method, would have been a good starting point for the investigation, this would not necessarily have led to understanding the variation in student thinking. Case study interviews are also time consuming and it remains unclear that a non-interview method will yield anything useful. Also, of interest was also identifying whether it was possible to design a non-interview methodology in order to investigate students' reasons for using the mean. Therefore, the present work comprised two small studies so as to get a flavor of the issues involved in trying to probe how students model things. The studies were carried out over a two-step developmental process. Study 1 was administered to 20 respondents. This was then increased to 30 respondents in Study 2. The present work reports on a structured approach on understanding meta-level questions using simple written instruments. i.e. to what extent can such an approach yield meaningful data that leads to meta-level questions. Below is a flow diagram highlighting the development of the questions used in Study 1 and Study 2. A more detailed schematic diagram is provided in Chapter 4.



Chapter 2: Study 1

As noted above the purpose of the present work was to use the PMQ or parts thereof as a reference instrument, owing to the fact that is well established, and then to modify or extend it in the areas of interest. In particular, to investigate how students think about data obtained from measurement at a more fine-grained level than that obtained from the Physics Measurement Questionnaire. The following section describes the methodology, analysis and main findings for Study 1, which was the first attempt at this exercise.

2.1 Development of the questions for S1

Four probes were selected from the PMQ for the development of the questions for Study 1: Repeating Distance (RD), Repeating Distance Again (RDA), Using Repeats (UR) and Straight Line Graph (SLG). The two *data collection* probes (RD and RDA) were selected as control questions while the *data processing* probes (UR and SLG) were the focus of this work and were adapted into four new questions. The adapted questionnaire (Instrument S1) consisted of six written questions as described below.

Previous studies have shown that both the posited context and question format have significant influences on how respondents answer the questions. This is related to issues around priming and framing. Thus, in order to be able to compare and relate the results from the pilot study to previous work, the new instrument was constructed so as to resemble the overall form of the PMQ. Hence, the same posited context and the first two questions of the PMQ (RD and RDA) was used as the opening questions of instrument S1. These control questions were also placed at the beginning of the instrument to ensure that the respondents were in the desired frame when they completed the subsequent reformulated questions.

UR was reformulated into three new questions. The structure of UR in the PMQ was open-ended with a calculational component which required students to calculate the final result of a set of repeated readings. While this question was able to produce good quality data, it was clear from the data analysis carried out in earlier work that there was still room for improvement and there were areas which could be investigated further. In particular, post-instruction questionnaire results showed that while many students were able to identify and calculate the mean as the best approximation, it was unclear why they carried this out. This therefore required further explanation and subsequently it was decided that it would be a good idea to explore the data processing probes further. Hence, the UR probe served as the starting point for the development of the questions for Study 1. Results obtained from the previous studies were also used in the construction of the questions.

UR was reformulated into three new questions so as to gain insight into the modelling approach students are using when calculating a mean. As found in previous work, the mean or average made up the largest emergent category for the UR probe after instruction. The recurring value was identified as the second largest emergent category. The adapted questions were not open ended like the ones in the PMQ but instead a debate format was used to probe these two categories further. This is described in more detail in the following section.

Question three (Using Repeated Distance with Explanation (UR1X)) investigated the respondents' thought process in using the mean as the final result in an experiment. Question four (Using Repeated Distance for Prediction (URP1X)) investigated the respondents' use of the mean to predict the outcomes of future data whilst question five (Using Repeated Distance in an Equation (URQ1X)) investigated the respondents' use of the mean in calculations. In each of the questions the mode or recurring value was introduced as an alternate measure of central tendency.

The final question (Straight Line Graph with Explanation (SLG1X)) looked at the respondents' decision process when combining data graphically. This question was adapted from the second data processing probe found in the PMQ (SLG). The original probe was open ended and required the respondents to draw the line of best fit for a given data set (said following a straight-line trend). SLG1X provided the straight line as the line of best fit and then probed the respondents about whether this was the right course of action. This question served to identify student ideas regarding the construction of the line of best fit and the data used to construct the line of best fit.

2.2 Framing the question

The pilot instrument followed the same structure and format as the Physics Measurement Questionnaire which was developed from a framework as described previously. All questions were framed in the form of a discussion where three or four posited students debated the nature of various laboratory procedures and data analysis techniques. The posited students were represented by genderless, raceless cartoons to avoid bias.

The respondents were required to select an opinion with which they most closely agree. The respondents were required to give a tick-a-box response followed by a qualitative description explaining why they made that choice. An example of what a typical question looked like for the instrument is given in figure 2.1.

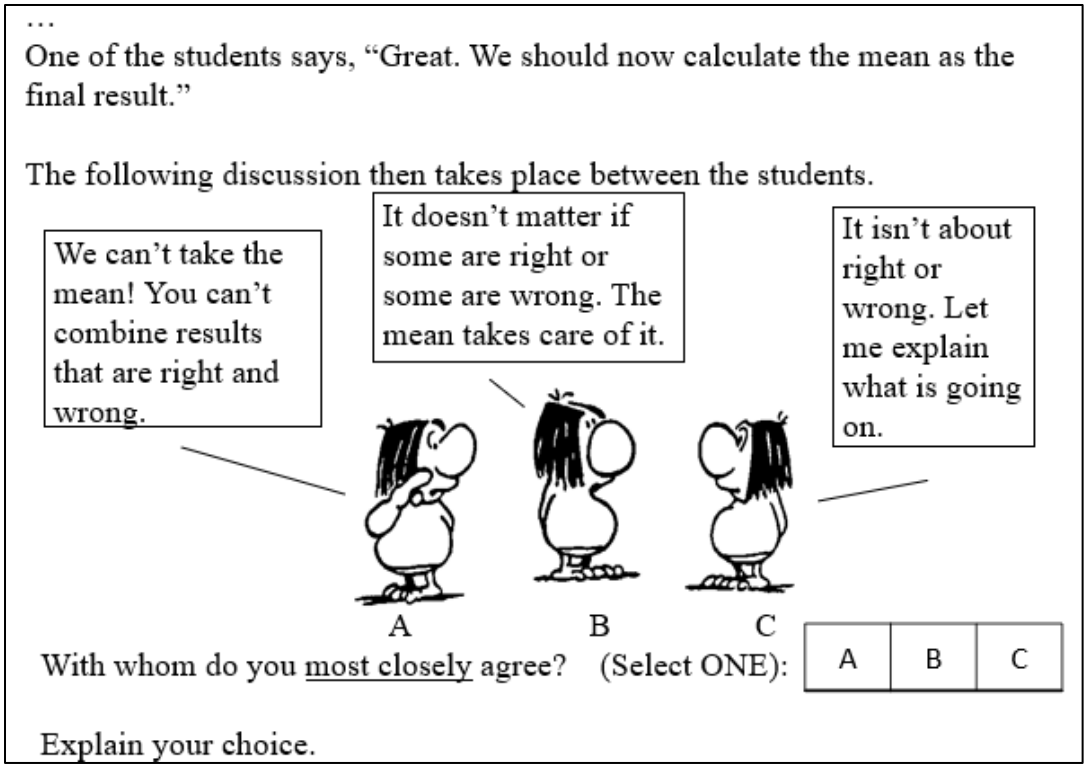


Figure 2.1: An example of what the structure and format looked like for the pilot instrument. The same structure and format was used that was developed for earlier studies.

Instrument S1 had six questions. These are given below.

Final questions constituting the instrument S1:

Below is a list of the questions in the order that they appear in the questionnaire. The full questions with the cartoons that accompanied each question can be found in Appendix 1.

Question 1 [RD]:

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

Question 2 [RDA]:

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	h = 400 mm = 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 <u>just one more time</u>.
C: Three releases will not be enough. We should release the ball <u>several more times</u>.

Question 3 [UR1X]:

The students continue to release the ball down the slope at a height h = 400 mm.
They obtain the following after six release:

<u>Release</u>	<u>d (mm)</u>
1	436
2	425
3	440
4	425
5	434
6	425

One of the students says, "Great. We should now calculate the mean as the final result."

The following discussion then takes place between the students.
A: We can't take the mean! You can't combine results that are right and wrong.
B: It doesn't matter if some are right or some are wrong. The mean takes care of it.
C: It isn't about right or wrong. Let me explain what is going on.

Question 4 [URP1X]:

The students take a bet on what result they will get for d if they release the ball again at a height h = 400 mm.

The following discussion then takes place between the students.

A: We have practiced a lot so I bet we will get the mean value.
B: No way! We will get the value that was repeated the most.
C: You are both wrong! Let me tell you what will really happen and why.

Question 5 [URQ1X]:

> The students then discuss what value to use for *d* in an equation.
>
> **A:** We have to use the mean value.
> **B:** We have to use the value that repeated itself the most.
> **C:** I don't think you understand what is going on. Let me explain.

Question 6 [SLG1X] (graphs excluded):

> For the next part of the experiment they measure *how the time to hit the ground changes if the* ball is released from *different heights*. They release the ball from 7 different heights *(h1, h2, h3... h7)* and measure the 7 corresponding times *(t1, t2 , t3...t7)*. They plot their data on a graph of *h* vs *t* as shown.
>
> One of the students says, "I drew a line through the data. We must forget about the points now and use this!"
>
> The following discussion then takes place between the students.
> **A:** Absolutely not! We have to use the data points exactly as we found them.
> **B:** No, in physics the straight line is what matters.
> **C:** I don't agree with any of you. Let me explain to you what we should be doing.

The full questions including diagrams can be found in the appendix.

Since the Using Repeats (UR) probe from the PMQ was purely calculational, the questions for this study were framed in a way that attempted to investigate the mean in particular and therefore was aimed at better eliciting their reasons for calculating a mean.

2.3 Administering the questionnaire

The questionnaire was administered to 20 non-majoring physics students at the end of the second semester in 2017. Therefore, these respondents would have attended and completed 8 practical sessions (laboratories) before completing the questionnaire. The practicals they completed were titled: Pendulum, Flow Rate, Spring, Turntable, Viscosity, Ball on Ramp, Flywheel and Vibrating String. The respondents were selected randomly. The respondents were informed that the data would be analyzed anonymously and that their lecturer would not be able to identify them by their responses. However, they were expected to provide a student number along with their questionnaire in the event that we might need to interview or question further respondents who provided interesting responses. This was also done to ensure some accountability was maintained by the respondents and that they provide meaningful and thoughtful responses. As a way to honor the anonymity promised to the respondents, after the data was collected the student numbers were replaced by Respondent Identification Numbers (RINs) and this was recorded in the analysis spreadsheets. The lecturer was not provided access to the raw data. The respondents were also informed that the questionnaire was being administered for research purposes. The

questionnaire was administered online using the Vula[6] system. The respondents were instructed not to discuss their answers with the other respondents. On average, the respondents took 15 minutes to complete the questionnaire.

2.4 Analysis

As stated previously, it was decided to pilot the instrument with a small number of respondents in order to be able to carry out a detailed analysis that would allow for a revised instrument to be tried out within the limited time frame of the present degree. The data was analyzed on a question by question basis. The analysis was broken down into two subsections:

- Analysis of Forced Choice Responses (FCR) (Quantitative Analysis)
- Analysis of Free Writing Responses (Qualitative Analysis)

2.4.1 Analysis of Forced Choice Responses (FCR)

The first step of the analysis of the FCR was to record the respondents' choice A, B or C for each of the probes. These can be found in Table 2.1. The columns are labelled according to the question number as well as an abbreviation describing each probe: Question 1: Repeating Distance Probe, Question 2: Repeating Distance Again Probe, Question 3: Using Repeated Distance Probe, Question 4: Using Repeated Distance for Prediction Probe, Question 5: Using Repeated Distance in an Equation Probe and Question 6: Straight Line Graph Probe.

Table 2.1: Forced Choice Responses: 6 questions x 20 respondents

RIN	Q1: RD	Q2: RDA	Q3: UR1X	Q4: URP1X	Q5: URQ1X	Q6: SLG1X
101	A	C	B	C	A	B
102	A	C	C	B	B	B
103	A	C	B	C	A	C
104	A	C	C	A	A	A
105	A	C	B	A	A	B
106	A	B	B	B	A	C
107	A	C	C	C	C	B
108	A	C	B	A	A	B
109	A	C	C	B	B	B
110	A	C	B	C	A	B
111	A	C	C	C	C	C
112	A	C	B	C	A	C
113	A	C	B	A	A	A
114	A	B	B	A	A	C
115	A	B	C	C	A	B
116	A	C	C	C	C	A

[6] This is UCT's in-house e-system for delivering course material to students. This was a departure from the way that previous PMQ exercises had been carried out at UCT (which had been via handwritten responses). However, there was a strong indication by the lecturer that the respondents had gained a good familiarity with computer usage and that the responses would not be compromised in any way.

117	A	B	B	C	A	C
118	A	C	B	C	A	B
119	A	C	B	B	A	B
120	A	C	C	C	C	C

Tallies of FCR data

For each of the probes the number of responses for each option was tallied to produce a bar graph for a graphical representation of the data. The number of respondents was reflected on the y-axis whilst the respondent's choice was reflected on the x-axis. These results are reflected in table 2.2.

Table 2.2: Tallies of options A, B, C for the FRC for each probe

Probe Question	Tallies: A	Tallies: B	Tallies: C
Q1	20	0	0
Q2	0	4	16
Q3	0	12	8
Q4	5	4	11
Q5	14	2	4
Q6	3	10	7

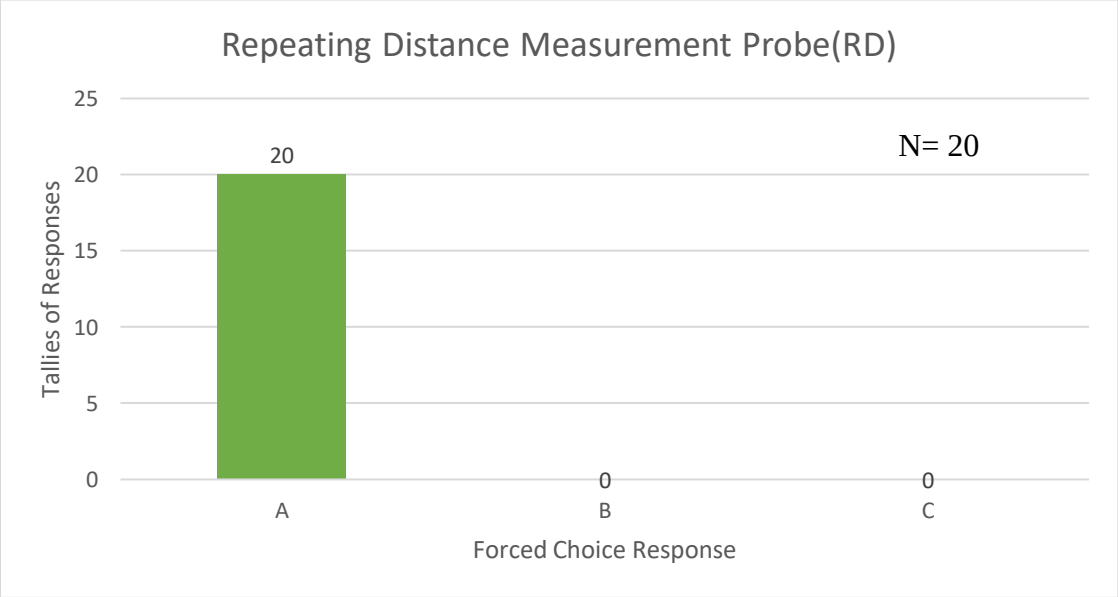


Figure 2.2: Histogram showing the distribution of forced choice responses for the repeating distance probe.

It can be seen in figure 2.2 that all of the respondents (100%) most closely agree with student A's view: "I think we should roll the ball a few more times from the same height and measure d each time." This option reflects an opinion which suggests that several repeated measurements from the same height are required for the rolling ball experiment. None of the respondents chose option B: "Why? We've got the result already. We do not need to do any more rolling." or

option C: “I think we should roll the ball down the slope just one more time from the same height.”

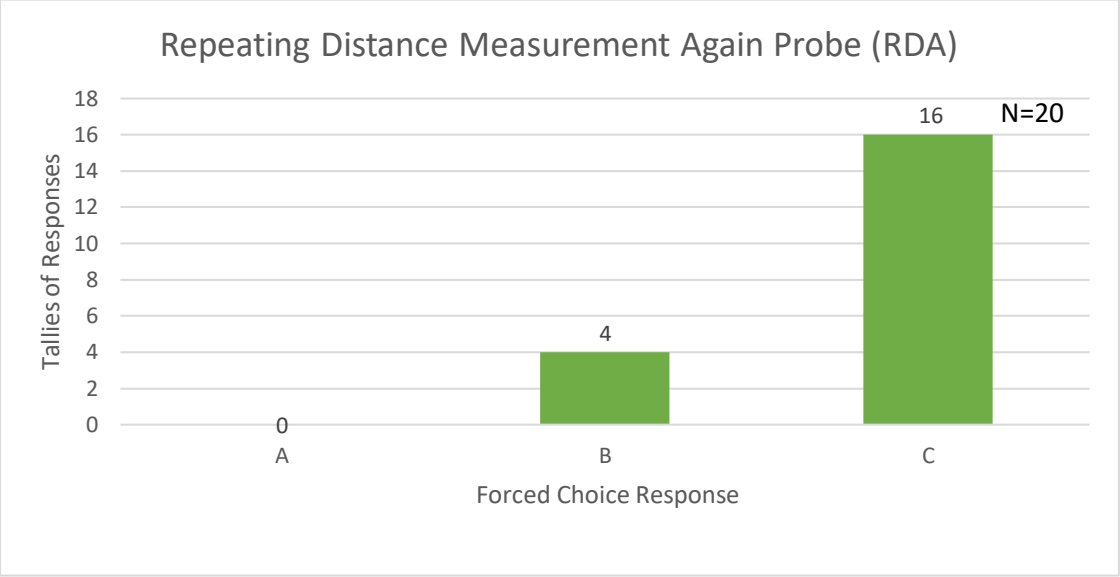


Figure 2.3: Histogram showing the distribution of forced choice responses for the repeating distance again probe.

For the repeating distance again probe, 80% of the respondents chose option C: “Three releases will not be enough. We should release the ball several more times.”, 20% chose option B: “We need to release the ball just one more time.”, while none of the respondents chose option A: “We know enough. We don’t need to repeat the measurement again.”

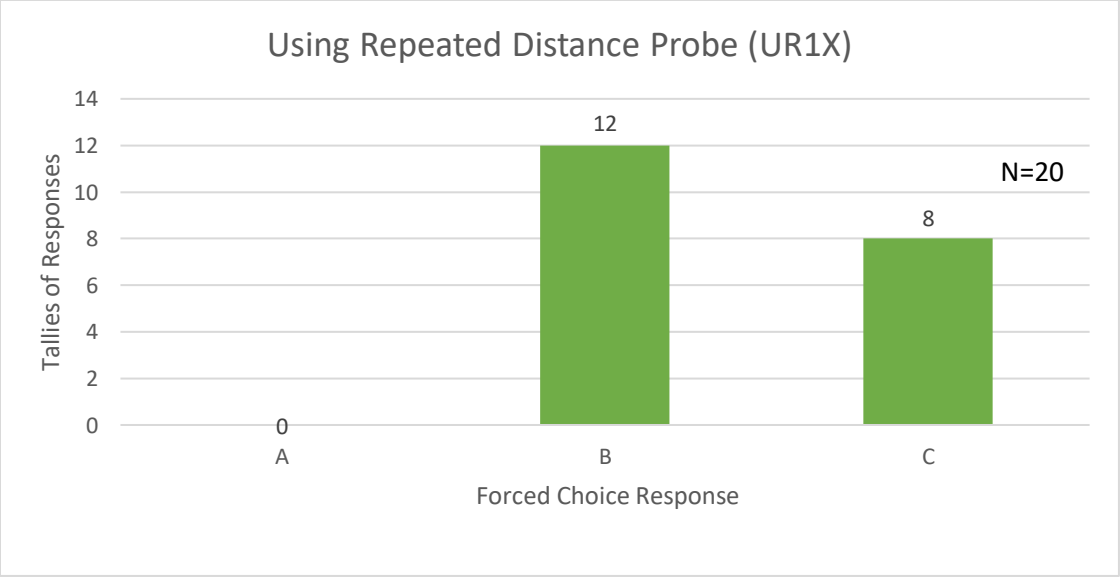


Figure 2.4: Histogram showing the distribution of forced choice responses for the using repeated distance probe.

Most of the respondents (60%) chose option B: “It doesn’t matter if some are right or some are wrong. The mean takes care of it.”. The remaining respondents (40%) chose option C: “It isn’t

about right or wrong. Let me explain what is going on.". While none of the respondents chose option A: "We can't take the mean! You can't combine results that are right and wrong.".

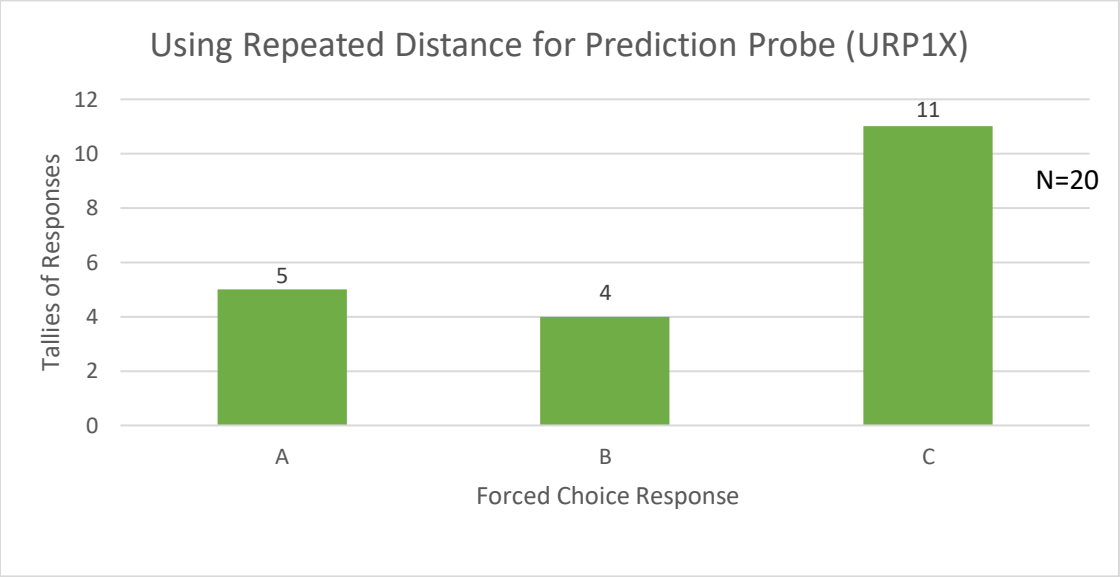


Figure 2.5: Histogram showing the distribution of forced choice responses for the using repeated distance for predication probe.

For the using repeated distance for prediction probe, 55% of the respondents chose option C: "You are both wrong! Let me tell you what will really happen and why." 25% of the respondents chose option A: "We have practiced a lot so I bet we will get the mean value." and 20% chose option B: "No way! We will get the value that was repeated the most."

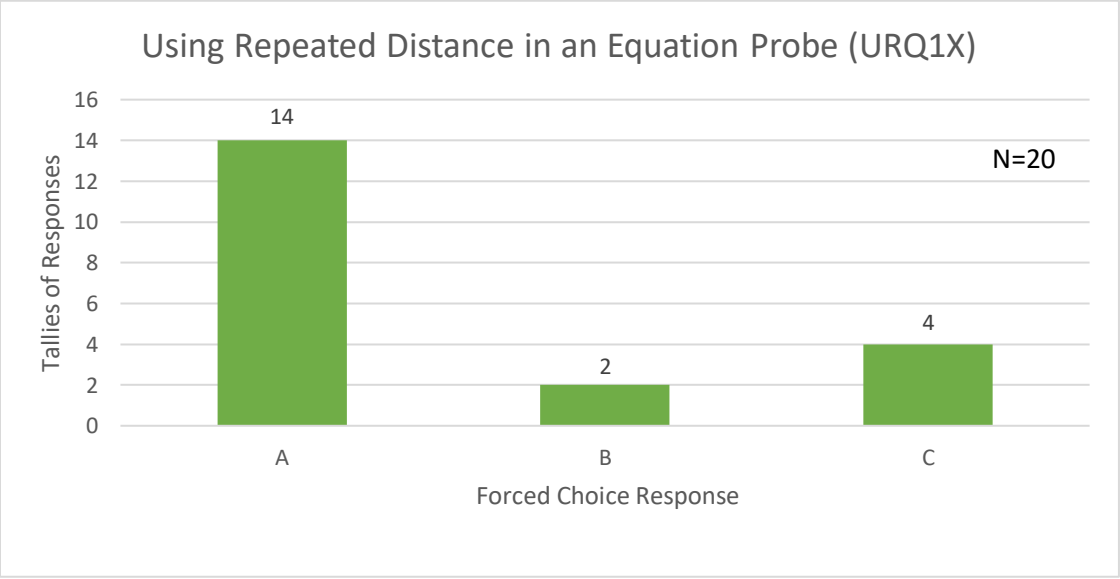


Figure 2.6: Histogram showing the distribution of forced choice responses for the using repeating distance in an equation probe.

For the using repeated distance in an equation probe (URQ1X), 70% of the respondents chose option A: “We have to use the mean value.”, 20% chose option C: “I don’t think you understand what is going on. Let me explain.”, while 10% chose option B: “We have to use the value that repeated itself the most. “.

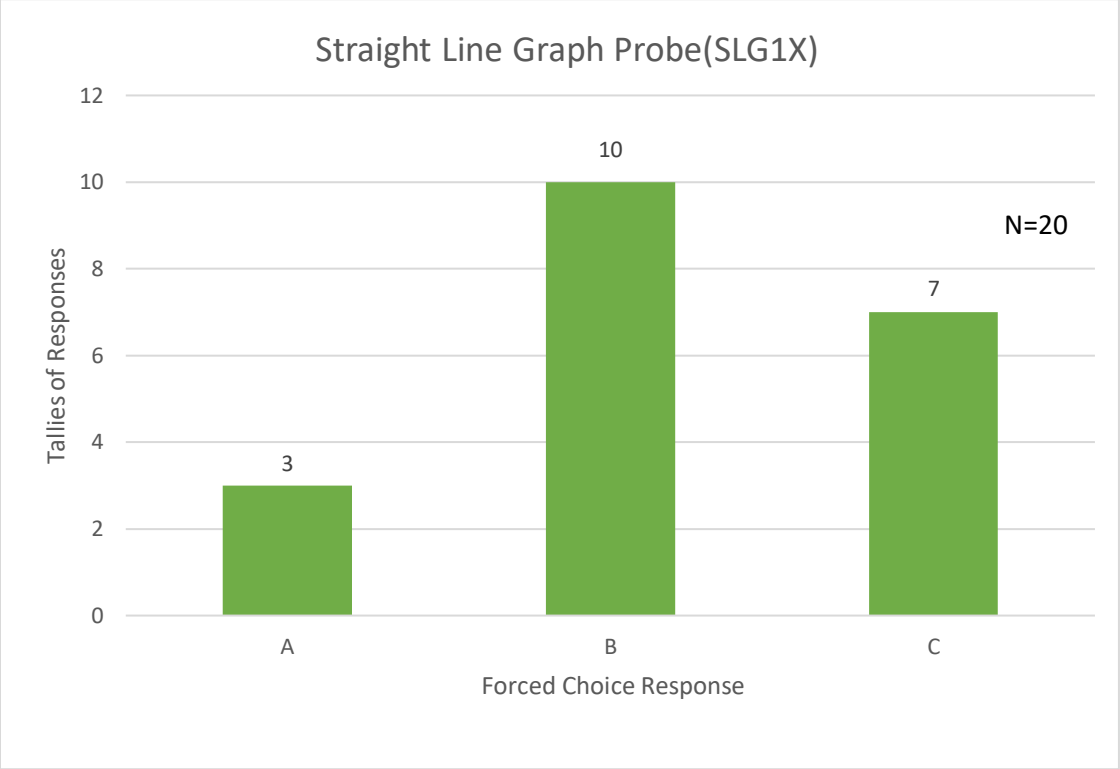


Figure 2.7: Histogram showing the distribution of forced choice responses for the straight line graph probe.

For the fitting a straight line probe, 50% of the respondents chose option B: “No, in physics the straight line is what matters.”, 35% chose option C: “I don’t agree with any of you. Let me explain to you what we should be doing.”, and 15% chose option A: “No, we have to use the data points exactly as we found them.”.

No analysis has been conducted across the FCR as we were more interested in the respondents’ reasoning for their choices i.e. in their Free Written Responses (FWR) for this study.

2.4.2 Analysis of Free Written Responses (FWR)

This section of this work aimed to explore the explanations for the respondents’ forced choice responses as provided by their free written responses. The analysis was done in two steps: (1) a Level of Informativeness ranking and (2) a themes analysis. The Level of Informativeness analysis framework is described in full below.

Methods suggested by Grounded Theory (Corbin 2008; Saldana 2009) served as a starting point for the analysis. However, as described in this chapter, the nature of the data did not allow for

emergent categories to be readily identified and therefore another preliminary analysis framework (ranking categorization) needed to be employed.

We encountered a similar situation in previous exploratory studies and therefore this analysis framework follows from our previous work (Bok, 2014; Majiet, 2016). In our previous exploratory studies, it was found that analysis of post-instruction questionnaires or questionnaires which investigate conceptually difficult areas of physics or astronomy can yield data that are not very informative i.e. inappropriate for a grounded analysis approach. This was either because the questions were above the students' level of expertise (for example, consider if you had to ask a high school student about something specific to special relativity their responses would not be based on anything "scientific" if they had not encountered that topic before) or they had not developed their own ideas about the topic. Therefore, a new preliminary analysis technique was piloted and used to determine whether the developed questionnaire was successful at yielding data of a high Level of Informativeness.

The first time this type of preliminary analysis technique was attempted was by Bok (2014) under the supervision of Professor Saalih Allie. As part of completion of her honours project, Bok established a methodology using a sophistication ranking to analyze a question from the Introductory Astronomy Questionnaire (IAQ) (Rajpal *et al*., 2014). For her study, the questionnaire was administered to students in an introductory astronomy course and the question she analyzed aimed to explore student ideas regarding the Big Bang as a theory for the starting point of the universe. The format and structure of the question was based on the same framework as the one used by us, i.e. the question was also framed in the form of a debate and required students to give a forced choice response as well as a free written response. Similar to our experience Bok (2014) first attempted to complete an analysis following a Grounded Approach and found the nature of responses to be unsatisfactory for this type of analysis. It was then agreed that it would be more useful to analyze the free written responses into 'levels of sophistication' as a simple measure to determine pre-post test shifts.

Table 2.3: Sophistication Levels (Bok, 2014)

Rank	Level of Sophistication
1	least sophisticated (LS)
2	intermediately sophisticated (IS)
3	most sophisticated (MS)

Due to the strong scientific nature and context of the question, as well as the aim of Bok's (2014) analysis (to determine pre- and post-test shifts) this level of sophistication ranking was correlated to the level of agreement with scientific supporting evidence.

The next time we attempted a preliminary analysis technique of this type was in 2016, in an honours project completed by me under the supervision of Professor Saalih Allie. In this project we explored student views on the relationship between theory and experiment. This question formed part of a broader study investigating student enjoyment of and engagement with physics laboratories (Tlowana, 2016; Majiet, 2016). After completing a Grounded Approach analysis on a sample of responses, we found that for this particular question, many respondents were unable

to fully elaborate on their responses and express their own ideas. We then decided that it would be more appropriate to carry out a sophistication ranking on this set of data. We found that many of the respondents were rephrasing statements provided in the question. After taking this into account we agreed that this response, while *seemingly* sophisticated, was not very informative about the respondents' level of understanding and so the criteria needed to be adapted to fit the context of this question.

Table 2.4: Sophistication levels (Majiet, 2016)

Rank	Level of Sophistication	Criteria
1	Least sophisticated (LS)	Restating the offered opinion
2	Intermediately sophisticated (IS)	Some form of explanation
3	Most sophisticated (MS)	A more reflective type of response

Translating this sophistication ranking over into our current work meant subjective inference was required by the researchers, since this ranking did not account for the positive (and perhaps false) bias created by a respondent's use of jargon in their responses. Therefore, such a response was likely to be ranked as having a high or intermediary level of sophistication. We felt that this was an unsatisfactory categorization especially when the aim of our analysis was to see the extent of student understanding about a particular concept and so the coding scheme was refined to the one below. We were then interested in the number of ideas expressed by the respondents and their ability to elaborate on these ideas. This was based on a paper by Allie *et al*. (2010) regarding the connection between student understanding and the Idea Space.

1 = Zero to one new idea presented,
2 = Two or three new ideas presented with little substantiation,
3 = Two or three new ideas presented with elaborate substantiation.

After receiving feedback on a peer reviewed paper for PERC Proceedings (Majiet, 2018), which used this coding criteria, it was agreed that the correlation between the ability of a student to express several ideas and sophistication of responses, was not entirely well established and that what we were rather interested in, was the **Level of Informativeness** of the responses. This proved to be a big step as it allowed us to realize that sophistication and informativeness each give us different information and the aim of the analysis will determine what type of categorization needs to be carried out. Since we were interested in the respondents' understanding of the mean and not necessarily in their ability to elaborate using scientific reasoning, we felt this to be a more appropriate preliminary analysis technique.

The way that one views the nature of science affects the way in which one views the nature of scientific measurement (Buffler *et al*., 2009). Therefore, in the context of measurement and uncertainty, what can be deemed to be "scientific" can range from one experimentalist to another. Therefore, to avoid researcher bias as much as possible, correlation or agreement with "scientific theory" was not used as a criterion when ranking student responses by level of informativeness. The categorization was mainly based on the clarity and elaboration of the ideas

expressed by the respondents. We chose to do this because we were interested in investigating student understanding of the mean as well student ideas regarding the mean after the completion of an introductory laboratory course.

The devised coding scheme used to rank responses according to level of informativeness is summarized in the table below:

Table 2.5: Level of Informativeness Ranking.

Level of Informativeness	**Criteria**
Low (LOI1)	No explanatory idea
Medium (LOI2)	Explanatory ideas with some elaboration
High (LOI3)	Explanatory ideas with considerable elaboration

It is difficult to assess the level of student understanding when vague explanations are provided and this requires a lot more inference from the researcher. Another shortcoming that we encounter when assessing post-instruction questionnaires is the use of jargon or technical terms by the respondents. In this instance it remains unclear whether the respondents have a good understanding of these terms or whether they merely recognize that the terms are relevant in answering the questions. Therefore, we argue that a response that has a high level of informativeness can include the use of jargon but there has to be further elaboration and clarity by the respondent to show that they do indeed have a good understanding of the terms they are using. We also argue that the ability to express more than one idea also provides a greater level of informativeness.

A non-explanatory idea reads like a statement of facts (often laced with jargon) with one or less main ideas (the respondent could just be repeating the question and therefore provides *no* new insight) while an explanatory idea has a main idea with supporting sentences or ideas. For example:

> An example of a non-explanatory response:
> *"The mean gives the average of the readings."*
>
> An example of an explanatory response:
> *"The mean gives the average; provides an approximation of the true landing distance; no right or wrong; does not have huge impact on the mean; unless most of the results are wrong in which you would get a false average"*

While more insight can be gained by asking multiple questions and then doing a cross probe analysis, we wanted to first assess how successful individual questions were at prompting the respondents to think more deeply.

2.4.2.1 Level of Informativeness Ranking

This section describes the Level of Informativeness analysis that was carried out on the free written responses. The first step in the analysis of the qualitative data was to read the free written responses and summarize the data without interpretation in a spreadsheet. The small sample of 20 free written responses was taken for each question and summarized in this regard. These were then recorded in the Summarized Written Response (SWR) column of the spreadsheet. The summarized written responses were then ranked according to the Level of Informativeness. Correctness with theory did not form part of the criteria when ranking responses. Therefore, the categorization was mainly based on the clarity and elaboration of the ideas expressed by the respondents.

Below are examples of responses for each category arranged according to the question which prompted the response. The responses were quoted without editing but in some instances inferred words were placed in brackets to provide clarity.

Repeating Distance Measurement Probe [RD]

A Level of Informativeness ranking was carried on the control questions to ensure that the respondents' level of engagement with this question was at a satisfactory level as was the case in previous work.

Table 2.6: Level of Informativeness Ranking for the Repeating Distance Measurement Probe.

Rank according to Level of Informativeness (RD)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=20)	5	9	6

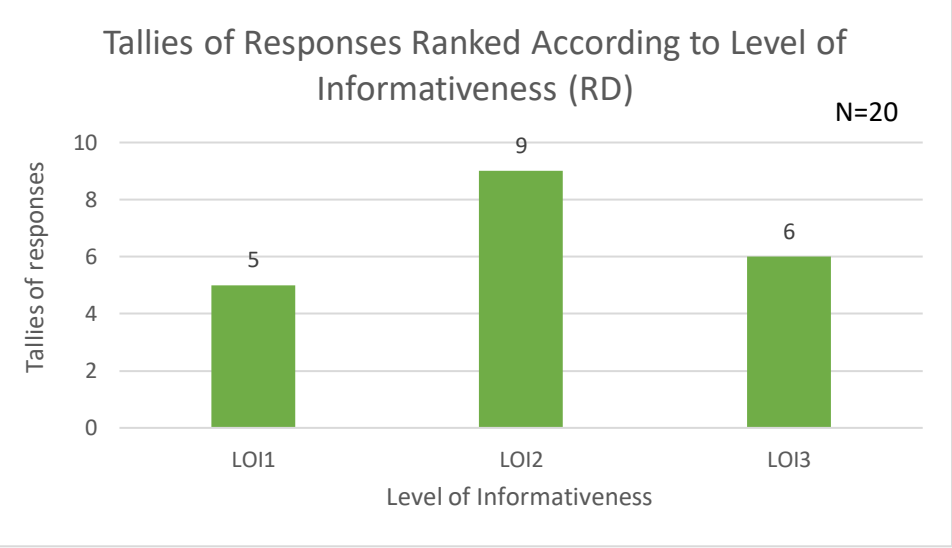


Figure 2.8:Histogram of the Level of Informativeness Ranking for the repeating distance measurement probe. The largest category is the level 2 responses.

An example of a level 1 response:
RIN 110: *"To get reliable results."*

An example of a level 2 response:
RIN 114 : *"To decrease uncertainty so that the results will be valid."*

An example of a level 3 response:
RIN 112: *"We know that if we roll the ball a couple of times it will not always land in the position it landed during the first trial, which means there is going to be an uncertainty in the experiment. So if we want to make the experiment more reliable we have to roll the ball a couple of times and take to average time."*

Repeating Distance Measurement Again Probe [RDA]

Table 2.7: Level of Informativeness Ranking for the Repeating Distance Measurement Again Probe.

Rank according to Level of Informativeness (RDA)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=20)	9	7	4

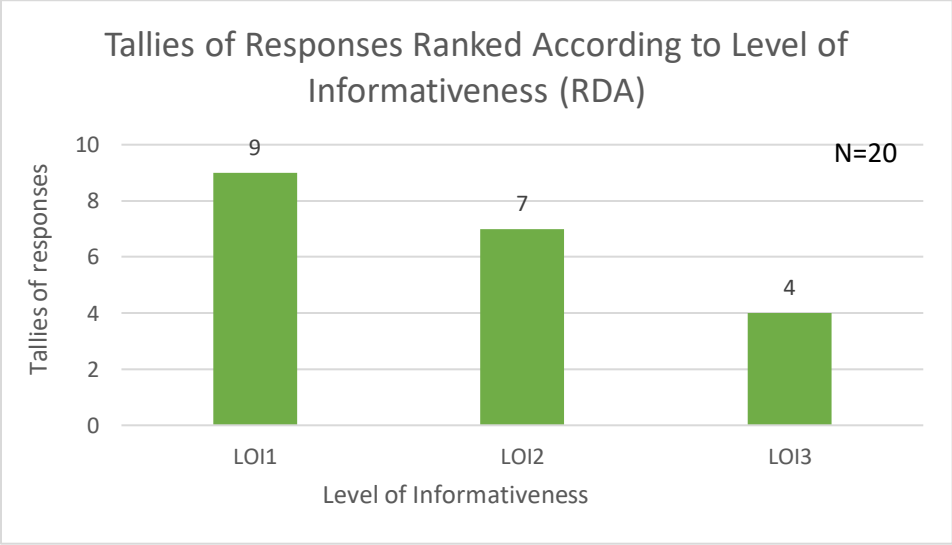


Figure 2.9: Histogram of the Level of Informativeness Ranking for the repeating distance measurement again probe. The largest category is the level 1 responses.

An example of a level 1 response:
RIN 102: *"more data values decrease uncertainty"*

An example of a level 2 response:
RIN 119: *"The two readings aren't in agreement with each other which is also a hint that more readings should be taken to improve the quality of results"*

An example of a level 3 response:
RIN 115: *"Since the measured distances are 10mm apart it shows that the true value is within 10mm so if a measurement is done one more time than the average of these 3 measurements will be ok. Taking a lot of measurements in this case will be a waste of time since the measurements are already closer to each other."*

Using Repeated Distance Probe [UR1X]

Question 3 was the first question that varied from the PMQ. It investigated the respondents' views of data processing, in particular the use of the mean as the final result in an experiment as well as the interpretation of variation in a repeated measurement experiment. The variation in a data set was framed as "right or wrong" results.

Table 2.8: Level of Informativeness Ranking for the using repeated distance probe.

Rank according to Level of Informativeness (UR1X)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=20)	8	8	4

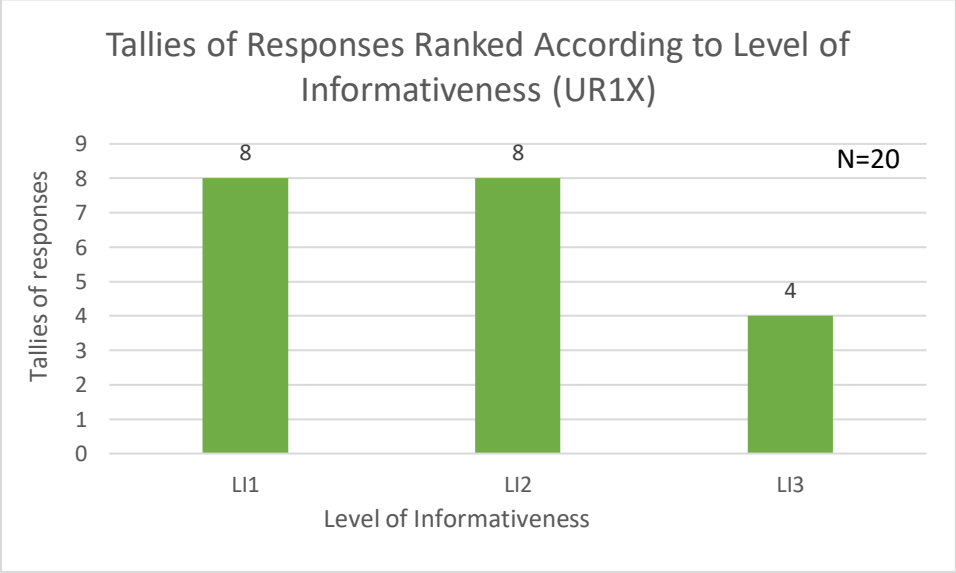


Figure 2.10: Histogram of the Level of Informativeness Ranking for the using repeated distance probe. The majority of responses were ranked as a level 2 or lower.

An example of a level 1 response:
RIN 110: *"The mean will give us an average value of all the readings."*

An example of a level 2 response:
RIN 112: "*The mean provides a better best approximation than the each measurement from the experiment that would have some uncertainty."*

An example of a level 3 response:

RIN 115*: "These results are due to the experiment not being in an isolated system, therefore many factors can affect the experiment and therefore yield different results – neither wrong nor right. Only an isolated system will yield exact results matching to theory."*

Predicting Repeated Distance Measurement for Prediction Probe [URP1X]

Question 4 investigated the respondents' views about predicting the outcome of future data after obtaining several repeated measurements. The question was framed in the context of a bet.

Table 2.9: Level of Informativeness Ranking for the using repeated distance for prediction probe.

Rank according to Level of Informativeness (URP1X)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=20)	10	6	4

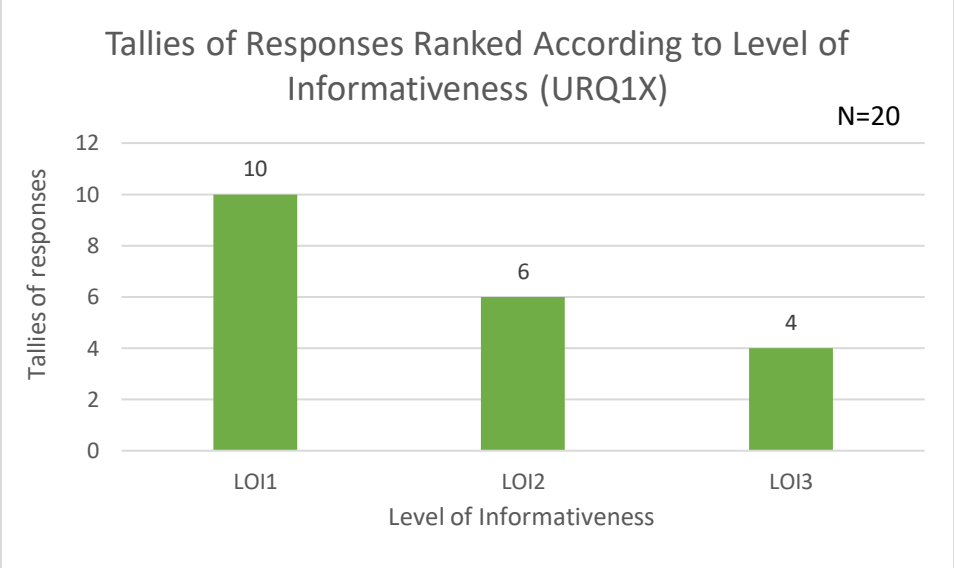


Figure 2.11: Histogram of the Level of Informativeness Ranking for the using repeated distance for prediction probe. The majority of responses were ranked as level 1 and 2.

An example of a level 1 response:
RIN 109: *"We will choose the value that appears the most."*

An example of a level 2 response:
RIN 110: *"Another reading will slightly change the mean value and definitely won't get the number that was mostly repeated."*

An example of a level 3 response:
RIN 111: *"The mean and the most achieved value are important, but they won't happen all the time. What will happen is that the ball should land somewhere within the uncertainty, or a few uncertainties of the mean."*

Using Repeated Distance Measurement in an Equation Probe [URQ1X]

Question 5 investigated student views on what to use of the final result of a data set to perform calculations.

Table 2.10: Level of Informativeness Ranking for the using repeated distance measurement probe.

Rank according to Level of Informativeness (URQ1X)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=20)	15	4	1

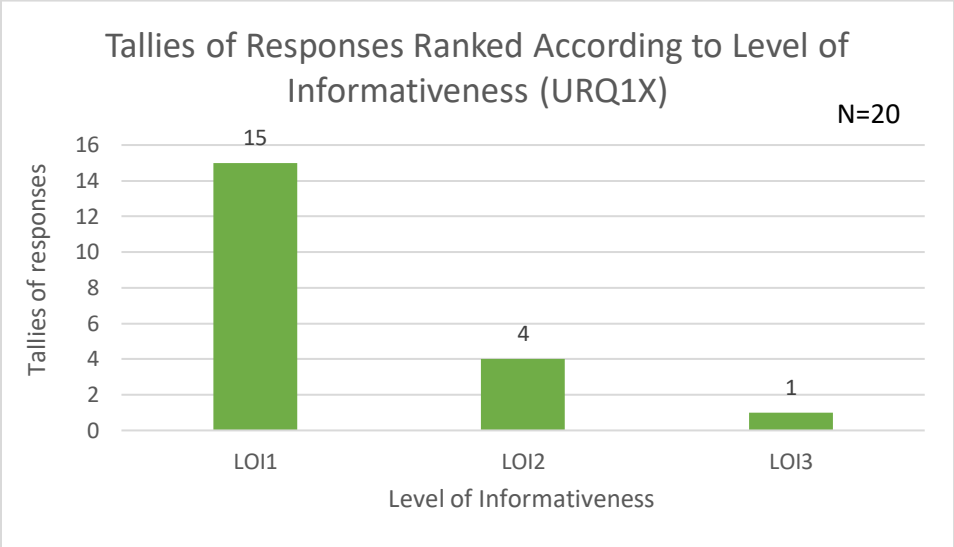


Figure 2.12: Histogram of the Level of Informativeness Ranking for the using repeated distance measurement probe. The majority of responses were ranked as a level 1.

An example of a level 1 response:
RIN 102: *"that value (mode) is very likely to be correct"*

An example of a level 2 response:
RIN 119:*"The mean value makes more sense since it represents widely spread data and even takes care of the odd readings (outliers)"*

An example of a level 3 response:
RIN 117: *"The mean value gives an approximation of the true distance in which the ball lands. Taking the mode or the repeated number would leave out other readings and would assume that the reading is the true distance landed while it has an uncertainty the same as the rest of the results collected."*

Fitting a Straight Line Probe (SLG1X)

Table 2.11: Level of Informativeness Ranking for the fitting a straight line probe.

Rank according to Level of Informativeness (SLG1X)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=20)	11	8	1

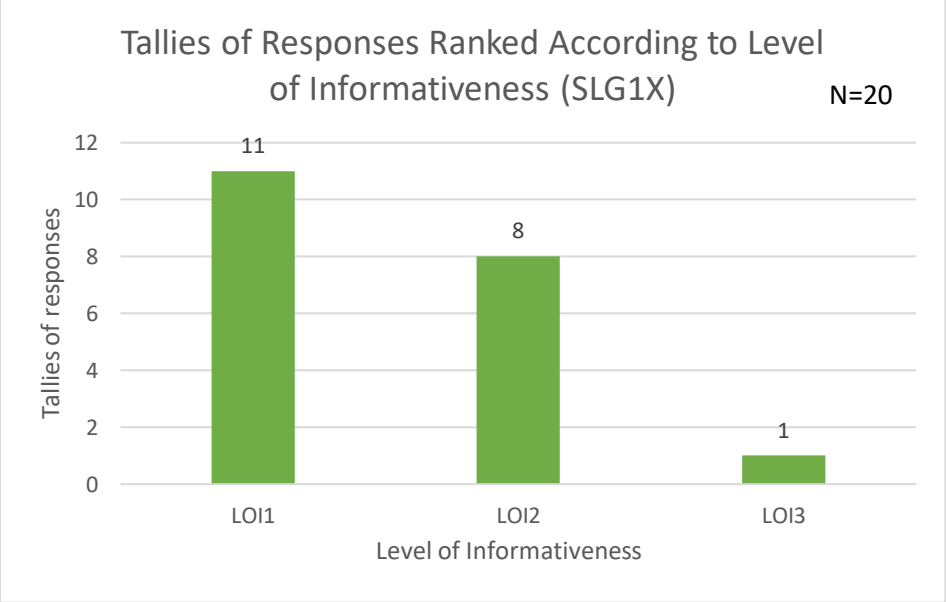


Figure 2.13: Histogram of the Level of Informativeness Ranking for the fitting a straight line probe. The majority of responses were ranked as a level 1 or 2.

An example of a level 1 response:
RIN 102: *"Yes since it is the line of best fit"*

An example of a level 2 response:
RIN 106: *"We use both the points and the line. the points will show the values we used and the line will show how close to the true value we are."*

An example of a level 3 response:
RIN 117: *"In an experiment we gather all results whether good or bad reliable or unreliable to plot a graph as we determine the relationship between the variables in this case it's distance and time. Then draw a line of best fit in which approximately determines the relationship"*

Table 2.12: Table of the tallies of responses according to the level of informativeness for each probe.

Level of Informativeness	Tallies of Responses					
	Q1 RD	Q2 RDA	Q3 MFR	Q4 PR	Q5 UR2	Q6 SLG2
Low (LOI1)	5	9	8	10	15	11
Med (LOI2)	9	7	8	6	4	8
High (LOI3)	6	4	4	4	1	1
Total	20	20	20	20	20	20

Table 2.13: Level of Informativeness across probes for each respondent.

Level of Informativeness across respondents						
RIN	Q1 RD	Q2 RDA	Q3 MFR	Q4 PR	Q5 UR2	Q6 SLG2
101	2	1	1	2	1	1
102	1	1	2	1	1	1
103	2	1	2	3	1	1
104	3	2	2	1	1	1
105	2	3	1	1	1	2
106	3	3	2	1	2	2
107	3	3	3	3	1	2
108	3	2	2	2	1	1
109	2	2	3	1	1	2
110	1	1	1	2	1	1
111	3	1	3	3	2	2
112	3	2	2	2	2	2
113	1	1	1	1	1	1
114	2	1	1	1	1	1
115	2	2	2	2	1	1
116	1	1	1	1	1	2
117	2	2	3	2	3	3
118	1	1	1	3	1	2
119	2	3	2	2	2	1
120	2	2	1	1	1	1

It is important to note that a response which was coded as “highly informative” was one which encompassed a large idea space (Allie *et al.*, 2010) and made it clear that the respondent was using their own ideas as opposed to one which would necessarily yield a conceptually “correct” response. To accommodate for this in the coding a plus or minus could be introduced, for example a 3- would indicate a response which is highly informative while a 3+ would indicate a

response which is highly informative and also conceptually coherent. Due to the small sample size and the nature of many of the responses this was not done in this pilot study.

It can be seen in all the examples provided above that the responses with a Low LOI had fewer words than those with Med LOI, which had fewer words than the High LOI responses. While it can be argued that this is a general trend with word count and LOI, we argue that it is possible to have both a large word count and a Low LOI. However, we did not find any strong examples of the contrary in our study i.e. a response that has a low word count and a High LOI. Typically, the High LOI responses had a large word count and the Low LOI responses had a low word count. However, the word count between the Low and Med LOI responses were comparable across questions. Below are some examples to highlight this. It should be noted that responses with a high word count are more difficult to analyze as the cohesiveness of ideas also tend to play a role in the interpretation of the response and its informativeness.

An example of a Low LOI response and Med LOI response with comparable word count:
RIN 120: "*No right or wrong; by calculating the mean you are decreasing the uncertainty*" (Low LOI)
RIN 119: "*No right or wrong; even if some values are a bit off; the mean counteracts that*" (Med LOI)

While both responses are difficult to understand and analyze, the first one uses more jargon than the other and the term *uncertainty* is used ambiguously. The second response also tells us something is going on in the respondent's head regarding the nature of the data points and is hinting at some deviation from an expected value i.e. the second response gives more insight into the respondent's own ideas.

Another example of a Low LOI response and Med LOI response with comparable word count:
RIN 105: "*The mean finds the best approximation to the spread out data; some results may be inaccurate*" (Low LOI)
RIN 102: "*Different factors will affect where the ball lands, so we don't really know what the right value is*" (Med LOI)

Once again the first respondent relies heavily on jargon while the second respondent does not.

Overall, what we found was that when students use more words, they are often attempting to motivate their ideas and fully explain their reasoning rather than string together multiple phrases containing mainly jargon. Therefore, there seems to be this general trend with word count and LOI. However, we do not want to restrict the LOI to be merely a word count as it is still possible for this to occur and therefore requires more judgment from the researcher.

2.5 Summary of findings: Study 1

The Level of Informativeness analysis showed that more than 50% of the responses were categorized as a level 2 or lower for each question. The majority of responses that were categorized as having a *high* Level of Informativeness came from the respondents who selected option C (the alternate option) for the FCR. Most of the level 1 responses consisted of definitions or jargon, for example: "*the mean is the center of location", "the mean is the most certain*

value", "the mean is more accurate" and *"the mean is the best approximation"*, while the level 2 and 3 responses were a combination of definitions and further explanations. Many of the responses were ambiguous and made it difficult to infer the level of student understanding of the nature of the mean.

The responses from the first two questions of Study 1's questionnaire indicated that more than 50% of the respondents used the expression "average" to motivate why they wanted to repeat measurements. Previous studies have shown that before instruction this term is used loosely by respondents and can have array of student interpretations (Allie *et al*., 2005). Findings for the present study indicate that the respondents still do not have a well-established understanding of the mean after instruction. A few case studies are highlighted below:

- *"Because it (the mean) is the most middle number and hence the most reliable one."*
- *"Using the mean provides a better best approximation because its in the middle of the interval and it (is) likely to be close to the actual value."*
- *"I'm not sure why we take a mean."*

It should be noted that while the first two responses quoted above may hold for a symmetrically distributed data set, the one provided was skewed and therefore a simple calculation made by the respondent would have shown them that the use of the mean as the best result and their motivation provided are not in agreement.

There also appears to be a limited understanding amongst respondents about the relationship between the mean (a discrete value) and the variation (represented by a distribution/interval) in a data set:

- *"The mean gives the average of the results, which is the value that we are most interested in. It shows how close our results are to each other.*
- *"The mean will give a rough estimate of the correct answer and if uncertainty is taken into account it will give a much closer value to the true value."*

Keeping this in mind the following responses are ambiguous to interpret.

- *"The mean value makes more sense since it represents widely spread data and even takes care of the odd readings (outliers)."*
- *"By calculating the mean, you are only decreasing the uncertainty."*

Question 5 provided the least informative responses for the questions regarding the mean. The majority of the respondents provided the same measure of central tendency for questions 3 and 5 while a few respondents switched from the mean to the mode when attempting to predict future data in question 4. For many of the respondents the responses they provided for questions 4 and

5 were also similar (word for word). The expression “the mean is the best approximation” was commonly used by the respondents throughout all the questions.

As the LOI analysis provides insight into the usefulness of the questions at eliciting respondents reasoning regarding their calculation of the mean, it was used as a way to inform the development of the questions in Study 2. Questions which had an overall low LOI was removed from the questionnaire. The LOI of the responses was also used as a guide to modify the questions for Study 2. For example, words (like “practicing”) that did not appear to affect the respondents’ reasoning were removed from the questions. Ideas which appeared often but had a low LOI were also used to modify the questions as we wanted to be able to investigate these ideas further. The findings described above were used to inform study 2 which is described below.

Chapter 3: Study 2

The questions from Study 1 were adapted based on Study 1's findings; these were then used in Study 2. The following section describes the changes that were made to the questions, methodology, analysis and main findings for Study 2. The set of questions for Study 2 will be referred to as Instrument S2.

3.1 Development and framing of the questions

The second version of questionnaire (Instrument S2) consisted of only four written questions. The overall structure and format of the questionnaire remained the same as the first version. Hence, the experimental context was kept the same and the control questions (RD and RDA) were placed at the beginning of the questionnaire.

Findings from Study 1 showed that the expression "the mean is the best approximation" was commonly used by the respondents. This was one of the types of reasoning that relied on ambiguous terms. Therefore, this formed a central part in the reformulation of question 3 (UR1X). A new question structure was piloted and replaced UR1X. The pilot question attempted to prompt the respondents to explain their responses even further. The piloted question structure can be seen in figure 3.1 and is explained in detail subsequently. The posited data set was changed so that it was identical to the one in the PMQ.

Question 4 (URP1X) was changed based on the responses from the first cohort but still followed the same structure and format as before. None of the respondents from the cohort commented on the idea that "practicing" a repeated measurement experiment affected the results therefore this was removed from question 4.

Question 5 (URQ1X) was removed from the questionnaire as it provided the least informative responses for the questions regarding the mean. The responses for URQ1X were also similar to the responses for URP1X (word for word) and therefore added noting useful for the purposes of the study.

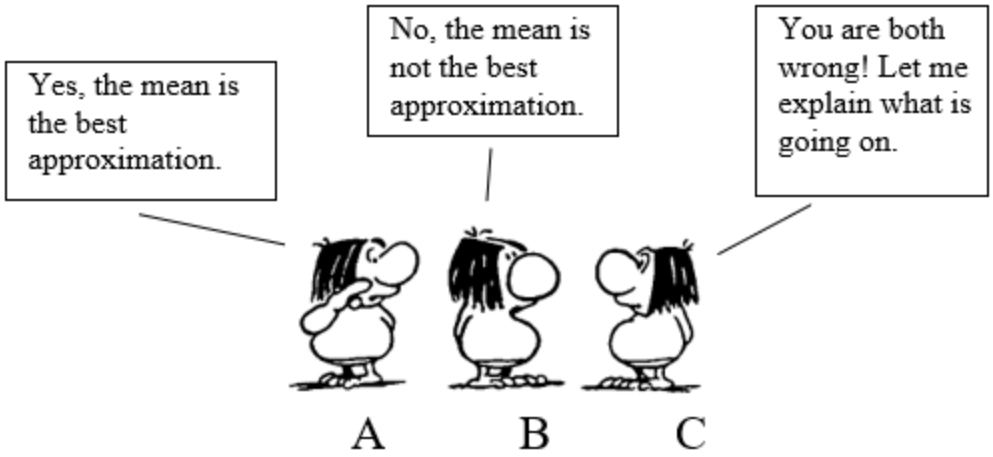
Question 6 (SLG1X) was removed from the questionnaire since the focus of the study narrowed from the broad area of *data processing* to *probing the mean*.

The students continue to release the ball down the slope at a height h = 400 mm.
They obtain the following after five releases:

Release	*d* (mm)
1	436
2	425
3	440
4	425
5	434

One of the students says, "Great. We can now calculate the mean as the result."

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

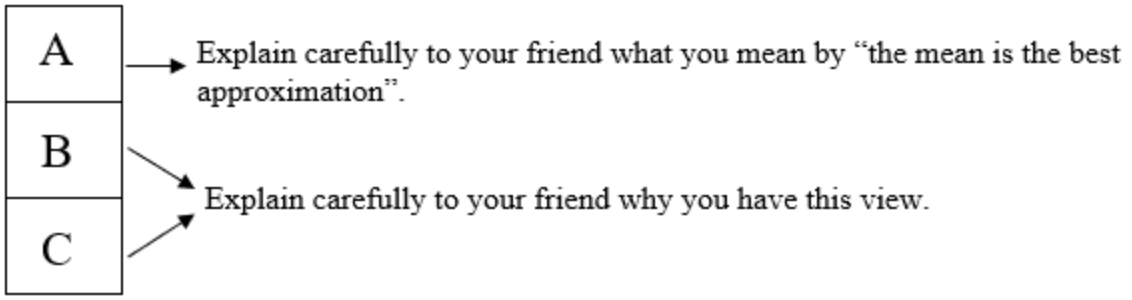


Figure 3.1: A new question structure was piloted. The options A, B and C were placed vertically and an arrow was then placed alongside each option. The arrow led to a different instruction based on the option chosen. The main idea was to get the respondents to use their own words and not rely on jargon to explain their reasoning.

For our study as well as the open-ended questions found in the PMQ, the majority of respondents identified the mean as the measure of central tendency to be used to describe the given data set (after having completed an introductory laboratory course). Therefore, UR2X for Instrument S2, option A stated that the mean is the best approximation, option B stated that the mean is not the best approximation while option C allowed for a different answer.

The positioning of the option selection was changed such that the options A, B and C were placed vertically in a column table rather than in a row. An arrow was then placed alongside each option. The arrow led to a different instruction based on the option chosen. The main idea was to get the respondents to use their own words and not rely on jargon to explain their reasoning.

Instrument S2 had four questions. These are given below:

Final questions constituting the instrument:

Below is a list of the questions in the order that they appear in the questionnaire. The cartoons have been omitted; the full questions can be found in the appendix.

Question 1 [RD]:

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

Question 2 [RDA]:

> 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 | h = 400 mm | d = 426 mm |
>
> The following discussion then takes place between the students.
>
> **A**: We know enough. We don't need to repeat the measurement again.
> **B**: We need to release the ball just one more time.
> **C**: Three releases will not be enough. We should release the ball several more times.

Question 3 [UR2X]:

> See figure 3.1.

Question 4 [URP2X]:

> The students take a bet on what result they will get for d if they release the ball again at a height h = 400 mm.

The following discussion then takes place between the students.

A: We will get the mean value.
B: No way! We will get the value that was repeated the most.
C: You are both wrong! Let me tell you what will really happen and why.

3.2 Administering the questionnaire

The new questionnaire was administered to 30 non-majoring physics students at the end of the second semester in 2018. These students completed the same laboratory course as the first cohort. Just as with the first questionnaire, the respondents were informed that data collected from the questionnaire would be used for research purposes. The respondents were informed that the data would be analyzed anonymously and that their lecturer would not be able to identify them by their responses. However, they were expected to provide an Emplid[7] along with their questionnaire in the event that we might need to interview or question further respondents who provided interesting responses. This was also done to ensure some accountability was maintained by the respondents and that they provide meaningful and thoughtful responses. As a way to honor the anonymity promised to the respondents, after the data was collected the Emplid were replaced by Respondent Identification Numbers (RINs) and this was recorded in the analysis spreadsheets. The lecturer was not provided access to the raw data.

Unlike with the first questionnaire, the second one was administered as a pencil and paper questionnaire so that the students could use diagrams in their explanations if they wished. The respondents were instructed to not discuss their answers with the other respondents.

3.3 Analysis

3.3.1 Analysis of Forced Choice Responses (FCR)

As with the previous data set, the first step of the analysis of the FCR was to record the respondents' choice A, B or C for each of the probes. These can be found in table 3.1. The columns are labelled according to the question number as well as an abbreviation describing each probe: Question 1: Repeating Distance Measurement Probe, Question 2: Repeating Distance Measurement Again Probe, Question 3: Using Repeated Distance Probe (with Explanation) 2.0 and Question 4: Using Repeated Distance for Prediction Probe 2.0.

Table 3.1: Forced Choice Responses: 4 questions x 30 respondents

RIN	Q1: RD	Q2: RDA	Q3: UR2X	Q4: URP2X
201	B	C	A	C
202	A	C	A	A
203	A	B	A	A
204	A	C	A	C
205	A	C	A	C

[7] An Emplid is a unique, seven-digit number issued to all students at the University of Cape Town.

206	A	C	A	C
207	A	C	A	C
208	A	C	A	A
209	A	C	A	A
210	A	C	A	C
211	A	C	A	B
212	A	C	A	C
213	A	C	A	C
214	A	C	A	B
215	A	C	C	C
216	A	C	A	C
217	A	C	C	C
218	A	C	C	C
219	A	C	A	C
220	A	C	A	A
221	A	C	A	B
222	A	C	A	A
223	A	C	A	B
224	A	C	A	C
225	A	C	A	C
226	A	C	C	C
227	A	B	C	C
228	A	B	A	C
229	A	C	C	C
230	A	C	A	A

Tallies of FCR data

For each of the probes the number of responses for each option was tallied to produce a bar graph for a graphical representation of the data. The number of respondents was reflected on the y-axis whilst the respondent's choice was reflected on the x-axis. These results are reflected in table 3.2.

Table 3.2: Tallies of options A, B, C for the FCR for each probe.

Probe Question	Tallies: A	Tallies: B	Tallies: C
Q1	29	1	0
Q2	0	3	27
Q3	24	0	6
Q4	7	4	19

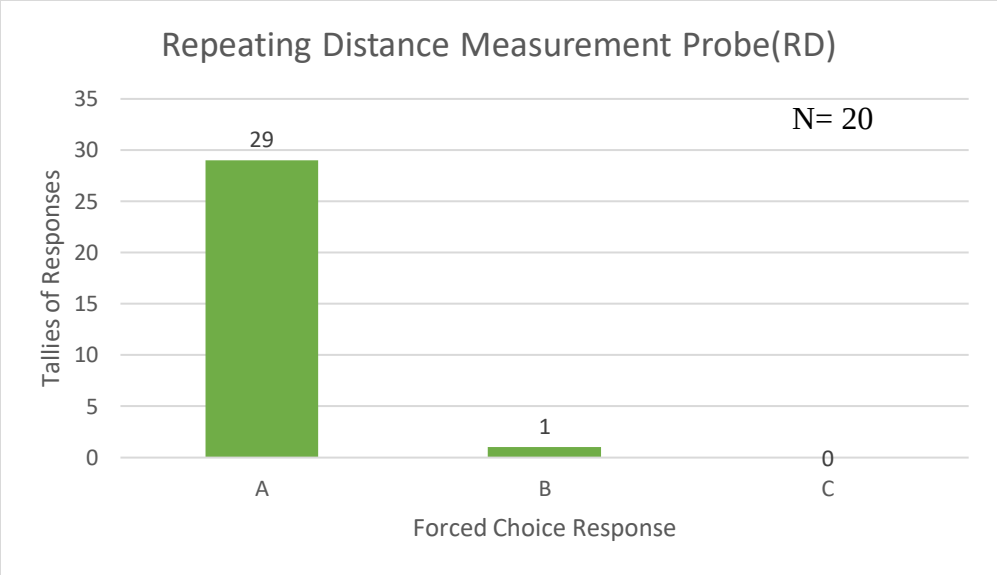


Figure 3.2: Histogram showing the distribution of forced choice responses for the repeating distance measurement probe.

It can be seen in Figure 3.2 that the majority of the respondents most closely agree with student A's view: "I think we should roll the ball a few more times from the same height and measure d each time." This option reflects an opinion which suggests that several repeated measurements from the same height are required for the rolling ball experiment. One respondent chose option B: "Why? We've got the result already. We do not need to do any more rolling." None of the respondents chose option C: "I think we should roll the ball down the slope just one more time from the same height."

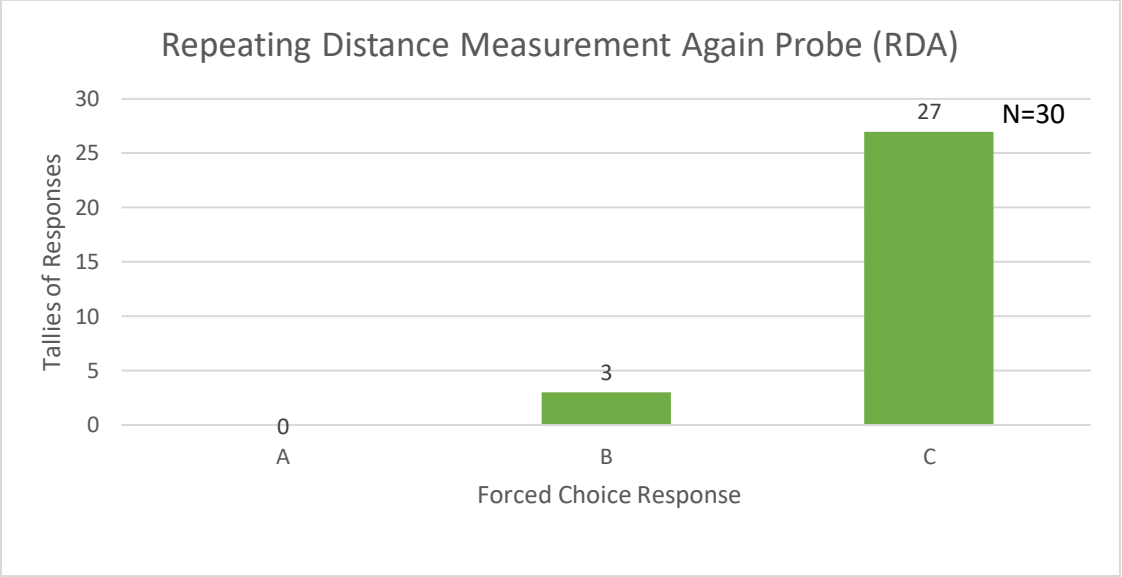


Figure 3.3: Histogram showing the distribution of forced choice responses for the repeating distance measurement again probe.

For the repeating distance measurement again probe, the majority of the respondents chose option C: “Three releases will not be enough. We should release the ball several more times.”. The remaining respondents chose option B: “We need to release the ball just one more time.”, while none of the respondents chose option A: “We know enough. We don’t need to repeat the measurement again.”

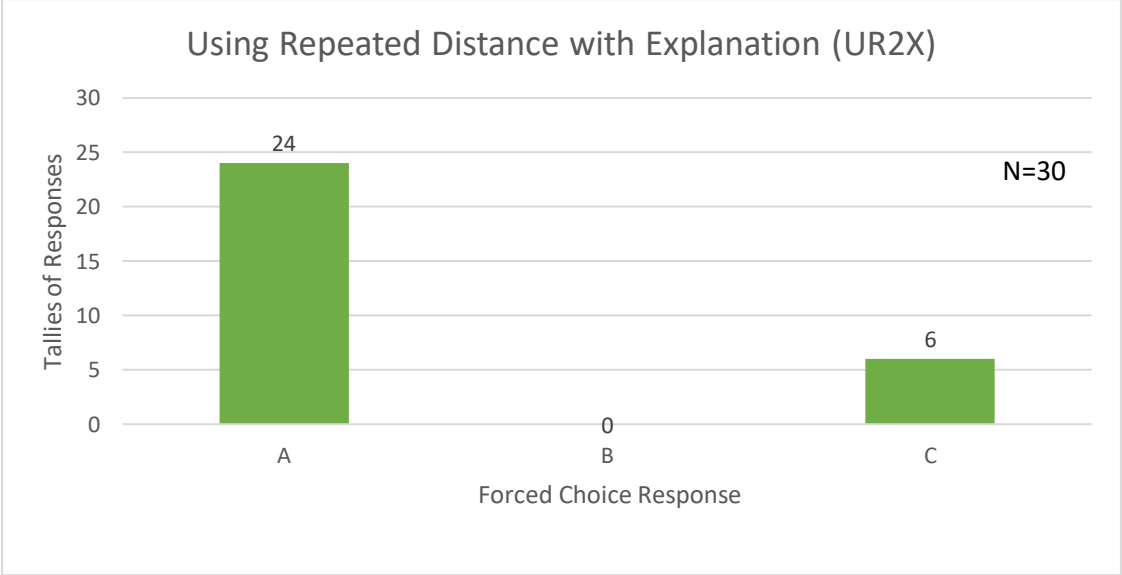


Figure 3.4: Histogram showing the distribution of forced choice responses for the using repeated distance probe.

The majority of the respondents chose option A: “Yes, the mean is the best approximation.”, while the remaining respondents chose option C: “You are both wrong! Let me explain what is

going on.". None of the respondents chose option B: "No, the mean is not the best approximation.".

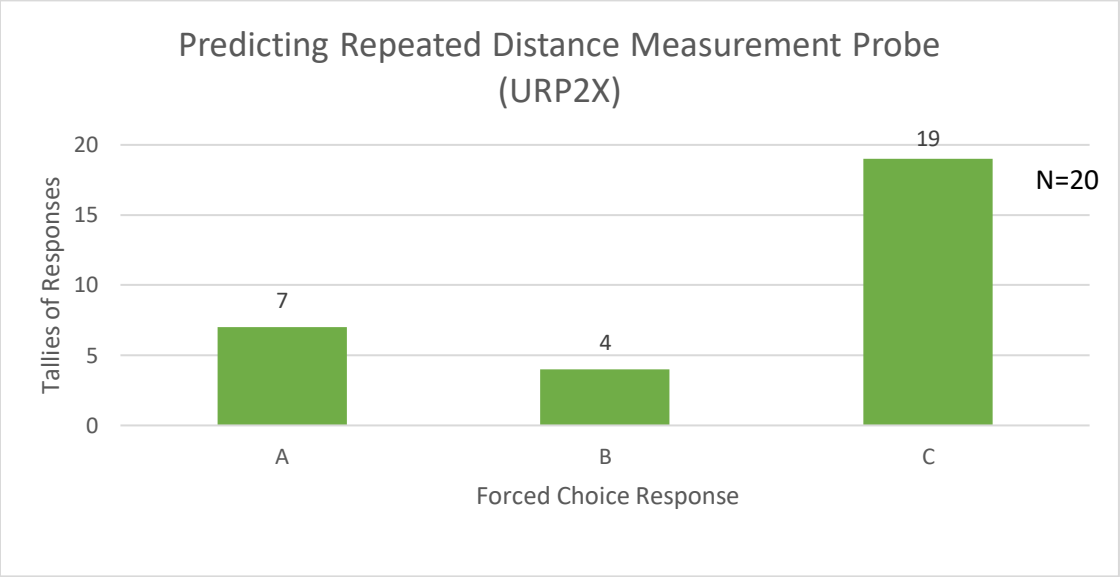


Figure 3.5: Histogram showing the distribution of forced choice responses for the predicting repeating distance measurement probe.

For the using repeated distance for prediction probe 2.0, 63% the respondents chose option C: "You are both wrong! Let me tell you what will really happen and why." 23% of the respondents chose option A: "We will get the mean value.", and 13% chose option B: "No! We will get the value that was repeated the most."

3.3.2 Analysis of Free Written Responses (FWR)

The analysis of the free written responses was divided up into two parts. The first part looked at the Level of Informativeness of the responses to determine the quality of the data and the extent to which the data could be analyzed using a Grounded Approach. The second part of the analysis looked at the ideas and themes that emerged from the responses.

3.3.2.1 Level of Informativeness ranking

The data for the new pilot questions were analyzed using the same analysis framework described in Phase 1. The small sample size of 30 free written responses were read and summarized without interpretation in a spreadsheet. These were then recorded in the Summarized Written Response (SWR) column of the spreadsheet. The summarized written responses were then ranked according to the Level of Informativeness.

Below are examples of responses for each Level of Informativeness category for questions 3 and 4. The responses were quoted without editing but in some instances inferred words were placed in brackets to provide clarity.

Using Repeated Distance Measurements with Explanations (UR2X)

Question 3 looked at the mean as the final result in a repeated measurement experiment. This question involved a discussion about whether the mean is the best approximation and what is meant by this.

Table 3.3: Level of Informativeness Ranking for the using repeated distance probe version 2.0.

Rank according to Level of Informativeness (UR2X)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=30)	8	20	2

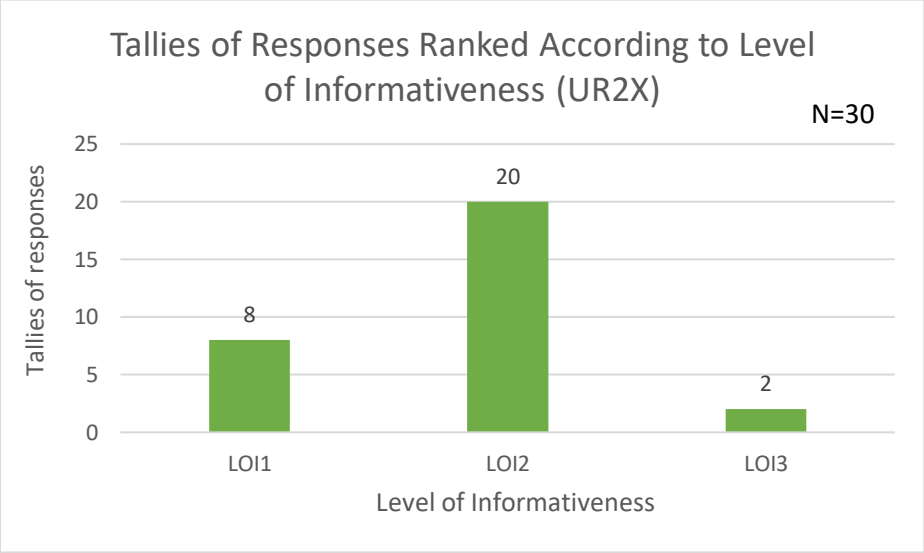


Figure 3.6: Histogram of the Level of Informativeness Ranking for the using repeated distance probe version 2.0. The majority of responses were ranked as a level 2.

An example of a level 1 response:
RIN 213: *"The mean will be the representive value of distance got/obtain by calculating the average of the distance"*

An example of a level 2 response:
RIN 207: "*Best approximation is more or less what the values of d will be. It's not the exact value but gives us a range of what/where the value might lie."*

An example of a level 3 response:
RIN 218*: "The mean is a good approximation of how close a result is the the true result, but an actual approximation is the difference in successive results (an uncertainty measurement). Therefore both the mean and the actual approximation (uncertainty) give a better reading d where the result lies so that a difference in results does not mean that the result is differentit may be different but falls within the uncertainty."*

Using repeated distance measurement [URP2X]

Question 4 investigated the respondents' views about predicting the outcome of future data after obtaining several repeated measurements.

Table 3.4: Level of Informativeness Ranking for the using repeated distance measurement probe version 2.0.

Rank according to Level of Informativeness (URP2X)	**LOI1**	**LOI2**	**LOI3**
Tallies of Choice (N=29)	6	15	8

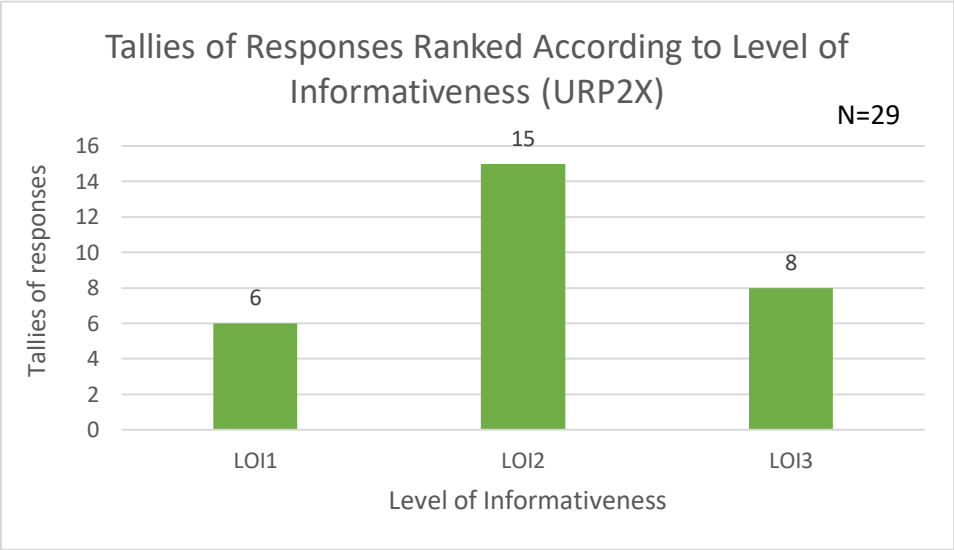


Figure 3.7: Histogram of the Level of Informativeness Ranking for the using repeated distance probe. The majority of responses were ranked as a level 2 or 3. There are only 29 responses as one of the respondents did not provide a FWR.

An example of a level 1 response:
RIN 202: *"You will calculate the average value for your data and use it to get reliable results"*

An example of a level 2 response:
RIN 228: *"The value will be within some interval due to some uncertainty. you cannot predict the exact value."*

An example of a level 3 response:
RIN 111: *"The value of the (next data point) cannot be what is repeated the most or the mean because some errors might occur during the experiment and the best way to get the approximation is to use uncertainty."*

3.3.2.2 Cross Probe Analysis

A cross probe analysis was then carried out on questions 3 and 4. This was done by comparing the responses for each question for each respondent. Each response was summarized so that the main idea(s) was identified. Table 3.5. summarizes what was found. Column 1 recorded the Respondent Identification Number (RIN), column 2 summarized the respondents' choice for the final result for question 3 and column 3 summarized the respondents' prediction in question 4. Three groups of respondents were identified: (1) those who were consistent in their responses across probes, (2) those who changed their responses from one discrete value to another and (3) those who chose a discrete value for question 3 and then identified that a range/interval or distribution was necessary to make a prediction in question 4.

An example pair of responses from group (1):

- UR2X Response for RIN 220:
  "By saying mean is the best approximation, it means we will get average d(mm) of the whole value and it will give us more precise or accurate answer, with less uncertainty"
- URP2X Response for RIN 220:
  "I think mean(average) is the better way of coming across this question because we have many values we need one value that will be more precise so we must calculate the mean meaning the average d(mm)."

An example pair of responses from group (2):

- UR2X Response for RIN 204:
  "It (mean) is the best approximation because it gives you the most likely possible value of the distance that the ball could roll up to. So it involves all values and approximates the general one."
- URP2X Response for RIN 204:
  "They might get any value as they roll the ball down, because the distance the ball can roll up to cannot be predicted. It occurs randomly at between any distance independent of what the mean is."

- *An example pair of responses from group (3):*
- UR2X Response for RIN 224:
  "The mean helps you find the average of all the measurements thus providing the more accurate answer"
- UR2X Response for RIN 224:
  "The height won't be the mean value but it will fall within the range of the mean value."

Table 3.5: Comparison of the respondents' choices for the final result and predicted value for URX2 and URPX2 respectively.

RIN	Question 3	Question 4
202	mean	mean
203	mean	mean
208	mean	mean
209	mean	mean
220	mean	mean
222	mean	mean
201	mean	neither
204	mean	any value
205	mean	any value
206	mean	different value
207	mean	different value
210	mean	other
211	mean	mode
213	mean	mode
214	mean	mode
216	mean	different value
219	mean	other
221	mean	mode
223	mean	mode
212	mean	number within range
224	mean	within range of mean
225	mean	within range of mean
228	mean	within some interval
230	mean	within range of mean
215	median	median
217	other	other
218	mean and uncertainty	value within uncertainty
226	mean and uncertainty	within range of mean
227	mean and uncertainty	use uncertainty
229	mean and uncertainty	within range of mean

3.4 Summary of findings: Study 2

Question 3 (UR2X) of study 2 looked at respondents' ideas regarding the mean; in particular, respondents were required to specify what they thought the best approximation was for a posited repeated measurement experiment. Question 4 (URP2X) then required the respondents to predict what the next data point would be if they were to repeat the experiment one more time. It was found that the most useful way to make sense of the responses was to analyze question 4's responses relative to that of question 3. For question 3, there were three types of responses: 24

(out of 30) respondents identified the mean as the best approximation, 4 identified the mean and uncertainty as the best approximation while 2 respondents argued that the most repeated value is the best approximation.

The subset of 24 respondents who identified the mean as the best approximation in question 3 was considered first. These respondents' responses for question 4 could be divided up into three categories; (1) The respondents who kept their responses the same for question 4 i.e. they selected the mean as their prediction for the next value, (2) the respondents who chose something else and (3) the respondents who predicted that the value will be within range of the mean value.

Table 3.6: This table shows the responses for the 6 respondents who motivated that the mean is the best approximation (for UR2X) and then predicted that if one were to take another reading, the mean would be the next value (for URP2X).

RIN	Question 3: Student ideas	Question 4: Student ideas
202	• The mean is the average. • You can use it instead of relying on one value that could be right or wrong. • The mean will give reliable results.	• The mean is the average. • The mean is reliable.
203	• The mean is the average. • The average is based on every roll put together.	• There was one value which was repeated. • Two values were very close. • These two values were not close to the repeated values. • The mean is more likely. • There is a broad range of data.
208	• It is representative of the measured values	• The mean is the middle value of a set of possible values
209	• The ball falls at different distances every time • we add up all of them and divide by the total number together	• It shows where the ball is likely to fall amongst the measured distances
220	• The mean is the average. • It will give a more precise or accurate answer, with less uncertainty.	• The mean is the average • We have many values • The mean is more precise
222	• The mean is the average	• No response provided

	• The mean is the distance that sums up all the distance measured	

This subset of responses was found to be uninformative and difficult to analyze. At first glance, the action (using the mean as the best approximation) can be categorized as a set action; however, the reasoning cannot be clearly identified as set reasoning. These respondents used jargon or technical terms (e.g. precise, accurate, reliable, etc.) to explain their choices. Previous work (Allie *et al*., 1998; Séré *et al*.,1993) showed that there exist many student difficulties with the usage of technical terms. Therefore, these terms cannot be used to identify student ideas regarding the mean. However, the presence of these terms might indicate that the respondents have fragmented ideas about the mean and therefore are using technical terms as a way to mask their conceptual difficulties. Of the ideas that are identifiable, there appears to be the use of point-like reasoning to explain their choice, for example, the point-wise comparison of data points.

Table 3.7: This table shows the responses for the 13 respondents who motivated that the mean is the best approximation (UR2X) and then predicted (URP2X) that the next measurement would be something other than the mean. They used a range of ideas to describe this (neither value, any value, different value, other value, most repeated value, etc.).

RIN	Question 3: Student ideas	Question 4: Student ideas
201	• The mean is the average. • We have lots of d which are different. • We need to add all the d's and divide by the number of d's.	• The last reading will just make the experiment more accurate.
205	• It is the most likely possible value. • It involves all values and approximates the general one.	• We might get any value. • It is independent of how many times we release the ball at the same height.
206	• The mean represents all values obtained • It is within range of almost all if not all the values obtained.	• They will get a different value. • The conditions at which the ball is released are not exactly the same for each release. • The values change. • The uncertainties also change.
207	• The mean is the average. • It takes into account every distance travelled by the ball.	• They will just get another value of what the value might be. • By adding this value to the others and dividing by the number of trials we can get the best approximation of d again.
210	• Best approximation is more or less what the values of d will be. • It is not the exact value.	• They will get the value that is close to the values they have.

	• It gives us a range of what/where the value might lie.	
211	• The mean lies within the range of values. • It is more accurate.	• The most repeated value has the highest probability of coming out. • The mean is an approximation.
213	• The mean will be the representative value of distance got/obtain by calculating the average of the distance	• d will be a random value. • It will be close to the most repeated value.
214	• The mean is the average. • It is the representative value.	• (Because it is a prediction) • The most likely value that they derived. • They derived this value from previous measured values. • It is most likely the most repeated value from previous calculations.
216	• It is within range of the highest and lowest values.	• They will get a different value. • Which is going to be used to calculate the mean. • This mean might be different if the value is different from the other values.
219	• You adding all the values. • It is most likely to appear.	• The mean will change. • The sum of releases and number of releases will change. • It does not depend on the number that is released the most.
221	• All values are around the mean value. • Can be used to estimate the values.	• The data revolves around the most repeated value.
223	• The mean is the average. • This mean lies between the smallest and biggest values.	• The most repeated value is most likely to be d.

This group of respondents had productive ideas about variation; however, they did not appear to appreciate the role of the mean in the characterization of variation. Some respondents mentioned the conditions of the experiment while others hinted towards the purpose of repeating the experiment and how this would influence the data set. For example, “The last reading will just make the experiment more accurate.”. Other respondents referred to how the mean would change if one were to take another measurement. 5 out of 13 of these respondents predicted the most repeated value as the next value. These respondents referred to the most repeated value as “most

likely" or "has the highest probability" even though for the previous question they motivated that the mean is the best approximation.

Table 3.8: This table shows the responses for the 5 respondents who motivated that the mean is the best approximation (UR2X) and then explained (URP2X) that that a range/interval/distribution around the mean could be used to predict where the ball might land. This was regarded as the ideal or most sophisticated response as ideas expressed in question 3 were used to make a prediction in question 4.

RIN	Question 3: Student ideas	Question 4: Student ideas
212	• The mean is the sum of all the values divided by the number of times you did the experiment. • The mean is the average. • The mean is inclusive of all the measured values.	• Will get a number within the range of numbers we already have. • d will be +- the same or similar to the other values. • We are rolling the ball from approximately the same height. • The value won't be too different.
224	• The mean helps you find the average. • It is the more accurate answer.	• The height won't be the mean value. • It will fall within range of the mean value.
225	• The average of the d divided by the number of releases. • It is more accurate.	• The value will fall within the uncertainty range of the mean value. • Not too different from the best approximation.
228	• We cannot get the exact answer.	• The value will be within some interval. • This is due to some uncertainty. • You cannot predict the exact value.
230	• The height was kept constant. • The mean is the average. • The general or likely distance is between the range of the lowest and highest value.	• They determined a mean value for the distance. • The ball will most likely end up somewhere in the region of the mean. • It will fall in this region with some uncertainty.

3 of these respondents explicitly mentioned uncertainty, however, it is important to note that the respondents used the term in different ways. For example, one respondent appeared to be referring to the uncertainty *of the prediction* rather than the uncertainty *of the data set* (which could be used to characterize the spread of data); *"The value will be within some interval due to some uncertainty."* while another respondent referred to "*the uncertainty range of the mean value".*

Table 3.9: This table shows the responses for the 4 respondents who motivated that the mean and uncertainty is the best approximation (UR2X) and used this to make a prediction in URP2X.

	Question 3: Student ideas	Question 4: Student ideas
226	• The mean is representative of all the values. • The mean is not the best approximation. • We need an uncertainty measurement as well.	• The value will be between the intervals of the mean and the standard deviation.
227	• An average must be determined and the uncertainty.	• The value cannot be what is repeated the most or the mean. • Some errors might occur during the experiment. • The best way to get the approximation is to use uncertainty.
229	• There will always be the same uncertainty to the instrument used. • You need to include the uncertainty to get the best approximation.	• They will get the value between the mean and uncertainty.
218	• The mean is a good approximation of how close a result is to the true result. • However, an actual approximation is the difference in successive results (an uncertainty measurement). • Therefore, both the mean and the actual approximation (uncertainty) gives a better reading. • A difference in results does not mean that the result is different. It may be different but falls within the uncertainty.	• They will get a result that falls within the uncertainty.

At first glance these respondents appear to be set reasoners, however the ideas appear to be fragmented. For example, there appears to be a conflation between the mean as the best approximation and the *final result* which is represented as the best approximation (often the mean) and associated uncertainty. While these respondents are using sophisticated terminology, it once again could be an indication of a fragmented understanding. It should be noted that Respondent 218, does not explicitly state that the mean and uncertainty is the best approximation and that this respondent might have a better understanding than the others in this category. However, this respondent does make mention of a "good approximation" and an "actual approximation" which indicates that the respondent has alternate ideas regarding the mean and uncertainty.

Table 3.10: This table shows the responses for the 2 respondents who selected something other than the mean as the best approximation for both questions. One respondent stated that the median is the best approximation while the other respondent identified the midpoint of a rectangle as the best approximation.

RIN	Question 3 response	Question 4 response
215	• "The best approximation is achieved using the median value."	• "When you throw the ball once more you get values from which you can calculate the median."
217	• "The best approximation, would be to put all these results in a rectangle and measure to the midpoint of that rectangle."	• "They will get a value close to the distance to the rectangle with all points in it."

These responses were not fully elaborated on and therefore difficult to analyze. The respondents also provided similar responses across questions, therefore comparison of responses across questions led to no new insights.

Chapter 4: Discussion

The purpose of the present work is to try and understand the extent to which students understand why they use the mean in order to represent a set of scattered readings. From the modelling perspective that was outlined earlier, this translates into trying to understand to what extent the calculation that is performed follows from an appropriate model or whether it is simply an arithmetic step that has no underlying basis. Seen from the PMQ perspective, a student who calculates the mean as an arithmetic step that has no underlying basis, would be a student who subscribed to the point paradigm but then used the mean as a calculational tool as opposed to a set thinker for whom the mean is a part of the set paradigm.

While the PMQ is useful for probing student understanding of aspects of measurement and uncertainty, it does not offer any insights into why students choose to calculate the mean as the question that involves the mean is purely action based. Thus, it leaves open the possibility that the procedure is carried out simply as a strategic procedure. While calculating the mean might be regarded as a step forward for students who were previously classified as point thinkers it can be argued that this is in fact a retrograde step from a modelling perspective in that the step can be described as "model abandonment". Studies based on the PMQ, for example, have indicated that students who started out as point thinkers and whose actions were consistent with this model were more likely to go on to becoming set thinkers than students who has previously started out calculating the mean as a rote procedure (Buffler *et al*., 2001). As stated in the paper:

"The present study has shown that it is possible for the actions and reasoning used by students in the laboratory to be drawn on an ad hoc basis from either the point or set paradigms, depending on the demands of the particular laboratory context. This is illustrated in figure 2 (see figure 4.1), where the four regions represent the four broad categories into which students may be classified based on both their actions and reasoning. Students whose reasoning and actions are both firmly rooted within the point paradigm may be located in the bottom left-hand region, while students who both act and reason according to the set paradigm may be located in the upper righthand region. These are the ideal cases of the point and set paradigms. Two other possibilities exist. Some students may be able to use the tools of statistical data analysis, i.e. are able to complete data analysis procedures associated with the set paradigm, but are theoretically rooted within the point paradigm. Such students therefore use the tools and actions of the set paradigm by rote. The fourth possibility in figure 2 is characterized by those students who have a coherent set paradigmatic view of measurement but who have not mastered the operational tools and procedures of data analysis. These students, therefore, use actions associated with the point paradigm…

"Our strong impression from teaching these students is that the procedural 'rules of thumb' acquired at school could be seriously impeding the development of procedural understanding at university. For example, students who join the data points on a graph when asked to 'fit' a straight line seem to be more easily introduced to the notion of a 'best fit' straight line than those who have come from school with an algorithm such as drawing a single line through as many points as possible. Furthermore, the notion of the mean as panacea for all the problems of experimental 'error', seems to impede the development of the ideas of inherent 'uncertainty' in measured quantities. It might be harder to shift students from the 'rote set actions' region of figure 2 to a coherent use of the set paradigm, than it is to shift students who both act and reason according to the point paradigm."

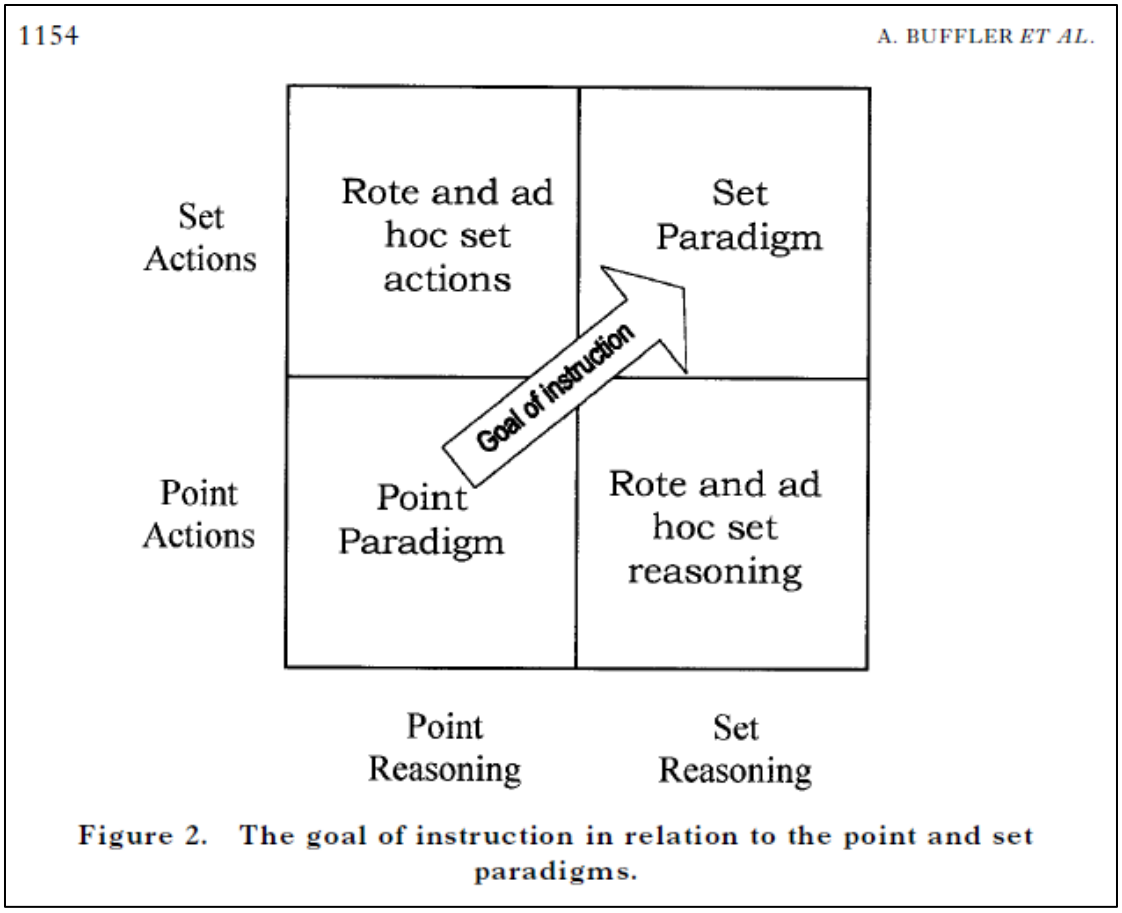


Figure 4.1: Figure taken from Buffler et al. (2001). The authors claimed that the goal of instruction should be to shift students from the point paradigm to the set paradigm. The authors showed that it might be harder to shift students from the 'rote set actions' region to a coherent use of the set paradigm, than it is to shift students who use consistent point reasoning and actions (bottom left region) to consistent use of set reasoning and actions (top right region).

Thus, it is clear that appreciating the way in which the mean forms part of a (set) modelling exercise is key to further understanding of deeper aspects of this model. However, as seen from the PMQ it is not easy to pose questions that probe what model, *if any*, students have in mind when calculating the mean. The present work thus explores to what extent it is possible to get some insight into the underlying thought process that accompanies calculating the mean. The formal study consisted of a two-stage process that tried to formulate suitable questions within the PMQ framework that would allow for some insight to be gained in this regard.

Stage 1 (Study 1) was a first (pilot) attempt at doing so while Stage 2 (Study 2) was a refinement based on experience gained from Study 1. Due to the timeframe of the master's degree it was decided to use two small samples (20 and 30 respectively) rather than individual case studies which would not have presented sufficient variation. The present work can thus be formulated in the spirit of "proof in principle" as it was not clear at the outset that any insight into the way in which students think about the mean would in fact be possible. The main stages of the present work are summarized schematically below.

4.1 Overview of the questionnaire development

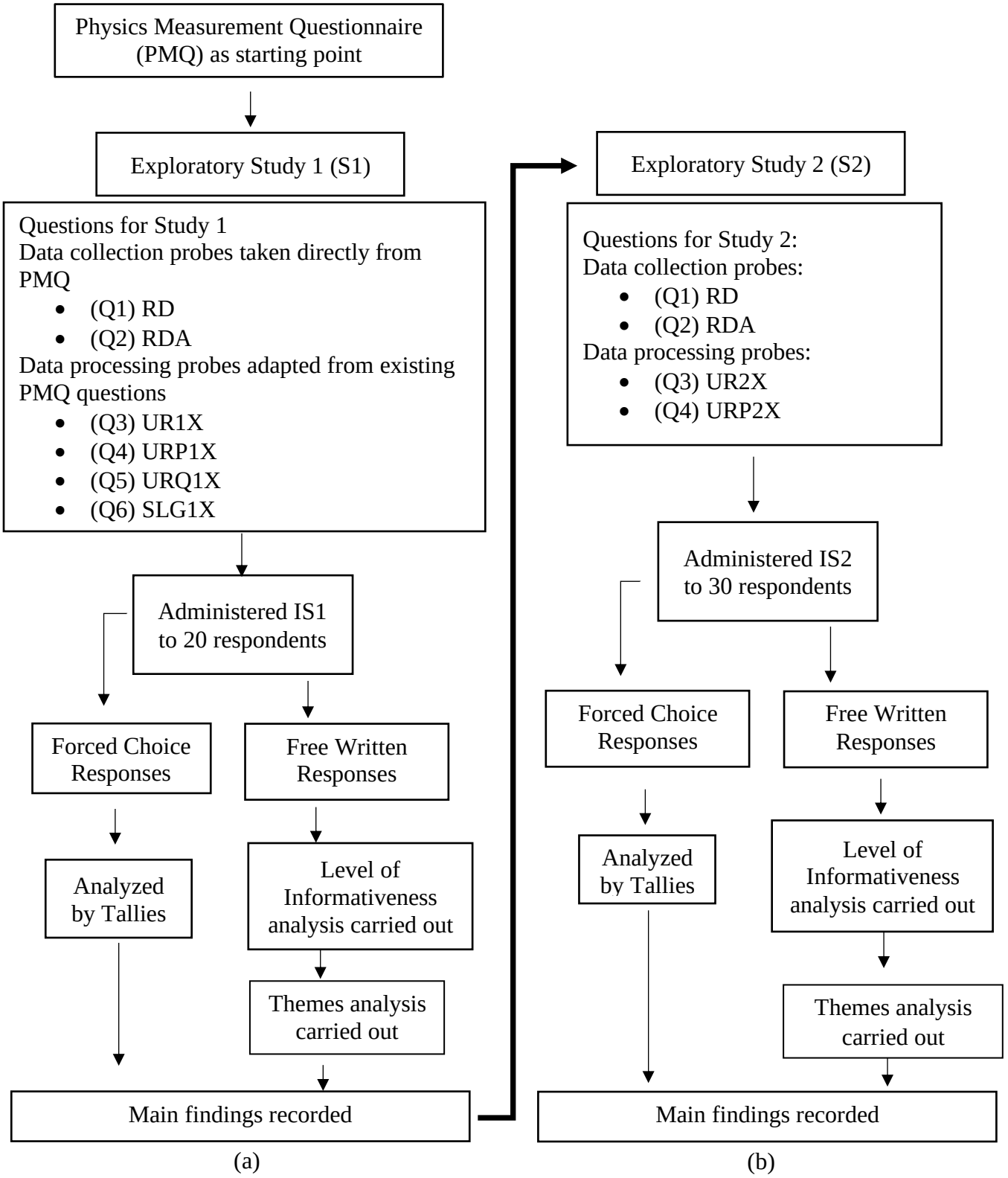


Figure 4.2: The present work consisted of a two-stage process ((a) and (b)) that tried to formulate suitable questions within the PMQ framework that would allow for some insight to be gained into student ideas regarding the mean.

Study 1 (fig 4.2(a)):

For Study 1, four probes were selected from the PMQ: RD, RDA, UR and SLG. The two data collection probes (RD and RDA) were selected as control questions while the data processing probes were the focus of this work. The control questions were put in place to ensure that the present respondents were in the same mindset as the PMQ respondents by the time they answered the data processing probes. This would then allow comparisons to be made with this work and earlier studies. The experimental context was also kept the same as the PMQ which followed a particular narrative with the questions placed in a deliberate order. The Using Repeated Distance Probe (UR) was an open-ended question with a calculational component in the PMQ. UR was adapted into three new questions all probing the decisions students make when calculating a mean. Each of these questions were framed in the form of a debate and the respondents had to choose the opinion with which they most closely agreed (Forced Choice Response (FCR)). This was to then be followed by an explanation for the choice (Free Written Response (FWR)). The first adapted question (UR1X) required the respondents to identify what to use as the final result from repeated measurements. The second question (URP1X) asked the respondents to make a prediction about future data while the third question (URQ1) asked the respondents to select what they would use in an equation. The questions were administered to 20 respondents at the completion of an introductory laboratory course. A Level of Informativeness analysis and cross probe analysis was then carried out on each of the probes. The responses for the RD, RDA, UR1X probes on average had a medium Level of Informativeness. Indicating that there was some level of engagement with the questions. However, for URP1X and URQ1X on average the responses were ranked as having a low Level of Informativeness. When the responses were compared across probes, it was found to be a useless activity as many of the responses were almost identical, particularly for URP1X and URQ1X. The findings from Study 1 were then used to inform Study 2 as described below.

Study 2 (fig 4.2(b)):

For Study 2 URQ1X and SLG1X were dropped from the questionnaire. URQ1X was dropped since the responses for URP1X and URQ1X were almost identical and therefore added nothing useful for the purposes of the study. SLG1X was dropped as the focus of the study narrowed from the broad area of *data processing* to *probing the mean*. Therefore, the second version of the questionnaire consisted of only four questions, two for control purposes (RD and RDA) and two questions probing the mean (UR2X and URP2X). UR2X and URP2X were adapted based on Study 1's findings. A new question structure as described earlier was piloted for UR2X in an attempt to elicit more meaningful responses. Unnecessary terms as found in Study 1 were removed from the URPX question. The questions were administered to 30 respondents at the completion of an introductory laboratory course. A Level of Informativeness analysis was carried on the probes and the majority of the responses were found to be of a *medium* Level of Informativeness. In particular, the overall Level of Informativeness increased for URP2X compared to that of URP1X, meaning that there was a higher level of engagement with the question for the second version of the questionnaire. A cross probe analysis was then carried out and gave rise to a medium Combined Level of Informativeness. This meant that when comparing the responses for UR2X and URP2X new insights emerged from the data. This is discussed in more detail later.

Table 4.1 below summarizes the substantive differences between the PMQ, Study 1 and Study 2 on a question by question basis as highlighted above.

Table 4.1: Probing the mean: Overview of the questionnaire development and Level of Informativeness analysis for Study 1

Studies	Probes	LOI	CLOI	
PMQ	RD RDA UR (Open-ended calculation) SLG	-	-	Mean calculation could be by rote, needs explanation
Study 1	RD RDA UR1X URP1X URQ1X SLG1X	2 2 2 1 1 1+	 - 1 -	-RD and RDA was taken directly from the PMQ for control purposes (student mindset same as for PMQ) -UR was developed into 3 new questions no longer open ended. -P1X = Q1X so Q1X seem to not be useful
Study 2	RD RDA UR2X URP2X	- - 2 2	 - 2	-Dropped URQ1X due to Low Level of Informativeness -Dropped SLG1X as the focus of study narrowed.

4.2 Usefulness of a Level of Informativeness analysis

The Level of Informativeness ranking is a preliminary analysis technique which allows the researcher to complete a quick analysis to determine whether a pilot questionnaire was successful at probing the relevant area of investigation. However, the main aim remains to be able to carry out a Grounded Approach analysis in order to identify what the emergent categories are. Whether a questionnaire is successful at probing the relevant area of investigation will be subject to what level of understanding the researcher is interested in exploring. Pre-instruction questionnaires typically aim to investigate respondents' intuitive and often framed as "naïve" ideas since instruction has not primed them, while post-instruction questionnaires typically aim to investigate the level at which respondents understand and are able to use certain concepts after instruction. Analysis of post-instruction questionnaires therefore are at risk of being prejudiced by seemingly sophisticated responses which are made up of jargon and technical terms. Therefore, to avoid this bias and fully investigate the respondents' ideas after instruction, the questionnaire needs to be able to allow the respondents to use their own words in their responses.

Since this study aimed to look at student understanding at a more fine-grained level than the PMQ, the criteria for what constitutes an informative response were stricter than was the case for the PMQ. It was found that the level of informativeness ranking was indeed a useful preliminary analysis to carry out when identifying if the pilot questionnaires were useful at providing insightful responses. This ranking allowed quick judgement to be made about the usefulness of the questionnaires even before attempting a Grounded Approach analysis. Therefore, it was possible to develop the second version of the questionnaire before extensive analysis was made on data unlikely to provide a fine-grained study of student ideas regarding the mean. The second

version of the questionnaire was more successful at probing student ideas regarding the mean. While the piloted question format was able to allow the respondents to elaborate more on their responses it also narrowed down what they chose to explain. This means that when designing a questionnaire using this framework, it has to be clearly defined what part of the response needs clarification. This cannot always be known until sample data has been collected. Therefore, the development of such a questionnaire needs to be developed over time to be as effective as possible.

4.3 Towards probing students' conceptual understanding of the mean

When comparing the responses for question 3 (UR2X) and question 4 (URP2X) new insights emerged from the data. Notably, there was a non-trivial Combined Level of Informativeness. The analysis showed that for UR2X, there were three types of responses: 24 (out of 30) respondents identified the mean as the best approximation, 4 identified the mean and uncertainty as the best approximation while 2 respondents argued that the most repeated value is the best approximation. The cross-probe analysis showed the greatest variation in student ideas came from the group of respondents who chose the mean only for UR2X.

The 24 respondents who identified the mean as the best approximation in question 3 could be divided up into three categories as described in figure 4.3 below.

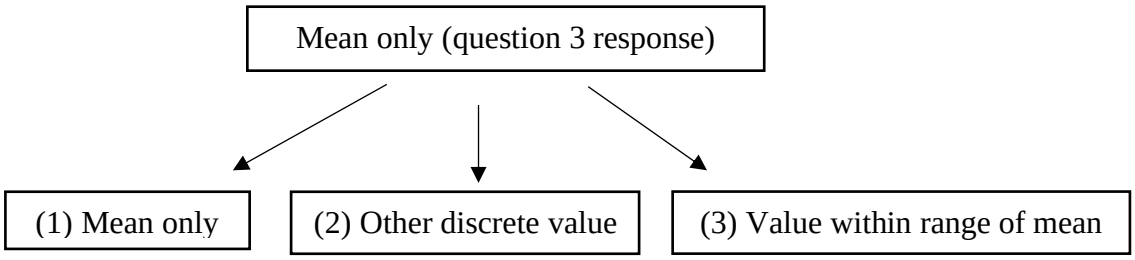


Figure 4.3: The group of students who selected the mean as the best approximation for UR2X can be divided into three groups based on their responses for URP2X: (1) The respondents who kept their responses the same for question 4 i.e. they selected the mean as their prediction for the next value, (2) the respondents who chose something else and (3) the respondents who predicted that the value will be within range of the mean value.

The group (2) pairs of responses made up about 40% of all the respondents. Their responses consisted of a range of different explanations. However, a common point was the lack of use of the mean to predict what potential values one could get despite motivating in question 3 why *the mean is the best approximation*. It can be argued that their answer for UR2X was more aligned with what they learnt in the lab course while their responses for the predictive question URP2X was more aligned with their pre-instruction intuitions i.e. with an everyday modelling approach.

In previous work students' responses were classified based on whether they were aligned with the point or set paradigm. This was done for each of the probes. The students were then classified within one of three categories: (1) consistent point reasoners, (2) mixed reasoners and (3) consistent set reasoners. This classification was based on the students' consistency across probes.

The range of fragmented ideas regarding the mean found suggests that the Using Repeated Distance Probe (UR), as it currently stands in the PMQ, might not be the most useful at determining whether students are point or set reasoners. Even though the majority of the respondents identified the mean as the best approximation for question 3, which can be categorized as a set action, this was often followed by an action which was more aligned with the point paradigm. Therefore, many of these respondents could be categorized as mixed reasoners. For many of the respondents, this only became apparent once question 4's responses were taken into account. While we suggest that the UR probe should remain part of the PMQ, we suggest that it should not be included as part of the analysis, in particular when determining if students are consistent point or set reasoners.

4.3.1 Addressing the key research questions

As outlined in section 1.6, the research questions for the study were:

(1) What reasons do respondents give when justifying the calculation of a mean?
(2) To what extent can we create a modified version of the PMQ that better elicits respondents' reasoning when calculating the mean of a set of data?
(3) To what extent does respondents' reasoning about means relate to other evidence for set or point like reasoning?"

We address the questions as shown above. It was shown in this chapter that respondents provided a variety of reasons to justify their calculation of the mean. However, many of these responses were difficulty to analyze and study further. It was difficult to create a modified version of the PMQ. The questions showed the difficulty respondents had with articulating their reasoning using their own words and ideas and this led to a low or medium level of informativeness with the majority of responses. It was not easy to relate the respondents' reasons to point and set reasoning. The responses were not clear enough and hence relating their reasoning to point or set like reasoning was shown to be challenging. As described in the previous subsection, the PMQ question alone can mislead the researcher into categorizing the calculation of the mean as a "set" move. More appropriately, the combination of the responses across the probes piloted in this study shows that these respondents are rather mixed reasoners. While the results remain inconclusive regarding whether this questionnaire better elicits respondents' reasoning when calculating a mean compared to the PMQ, the results do suggest that the open-ended PMQ question as is, does not provide enough information to allow the researcher to definitively categorize the reasoning as point or set like for that question alone.

4.4 Concluding remarks

Based on the approach detailed by Hestenes (1992), modelling lies at the heart of physics. In order to engage meaningfully with a physics task, it is necessary to be aware of the model that is being used. In theoretical physics (including first-year physics textbooks) there are many examples of students carrying out calculations without knowing the reasons for doing so. This applies even more strongly where experiment is concerned as it is often not even recognized that there is an underlying model behind the actions and calculations that are being carried out any one time. For example, the actions that are performed such as deciding on how many readings to

take or how to select or combine data into a representative result, only have meaning when they are understood in terms of the model that is being used.

From a learning perspective there is ample evidence to suggest that simply following recipes for problem solving, whether theory or experiment, has a negative effect on laying down a disciplinary foundation. The often-quoted phrase "What formula must I use?" comes to mind. While this phrase is more commonly associated with problem solving in first year physics it is equally evident in the laboratory where the data analysis process easily slips into an "applying formulae" approach.

Data analysis and uncertainty have proved to be challenging aspects of a first-year curriculum. This is so not because of the calculational aspects which are usually well mastered by novices but rather because the conceptual underpinnings are not well understood. The findings of studies carried out using the PMQ have indicated that students who start out with a point paradigm perspective appear to have successfully adopted a set paradigm (Pillay *et al.,* 2008*).* One of the markers of this is that students will calculate the mean rather than use the tools that are associated with the point paradigm (Volkwyn *et al.,* 2008). However, this step can easily be carried out *without any model in mind.* Thus, rather than the mean being a stepping stone to further understanding of uncertainty, it could in fact prevent such a learning trajectory. This is consistent with the UCT-York group who have pointed out that students who calculated the mean prior to instruction but otherwise were point paradigmers did not appear to benefit from the instruction. (Allie *et al.,* 2001*;* Buffler *et al.,* 2003*).* It is therefore clear that an understanding of how the mean fits into a modelling of data framework is an important transition point into deeper understanding of uncertainty.

The present work thus aimed to explore the degree to which it was possible to identify to what extent students used the mean with some model in mind. The starting point for the work was the PMQ which while it has proved to be useful at categorizing students broadly, does not probe this particular crucial aspect from a sense-making perspective.

The present studies that were carried out indicated that there was no straightforward way to elicit information as to whether the student had some model in mind or not insofar as deploying the mean was concerned. However, a number of insights into the way forward were gained. These included the way in which questions could be framed around the issues of the mean that allowed for some level of inference to be made. While further work still remains insofar as this is concerned, we suggest that such questions be included in future versions of the PMQ. Furthermore, ways of eliciting responses that indicate how data are being modelled need to be pursued so that sense-making underlies laboratory work and data reduction.

The laboratory offers one of the few places in undergraduate physics teaching that has the potential to develop critical thinking skills along the way. However, this requires that sense-making underlies each step, in particular with regard to data reduction that forms the basis of epistemic claims that can be discussed meaningfully by students.

Appendix 1: Study 1's Instrument: S1

EMPLID:

S1

University of Cape Town
Department of Physics

Laboratory Questionnaire

Instructions:

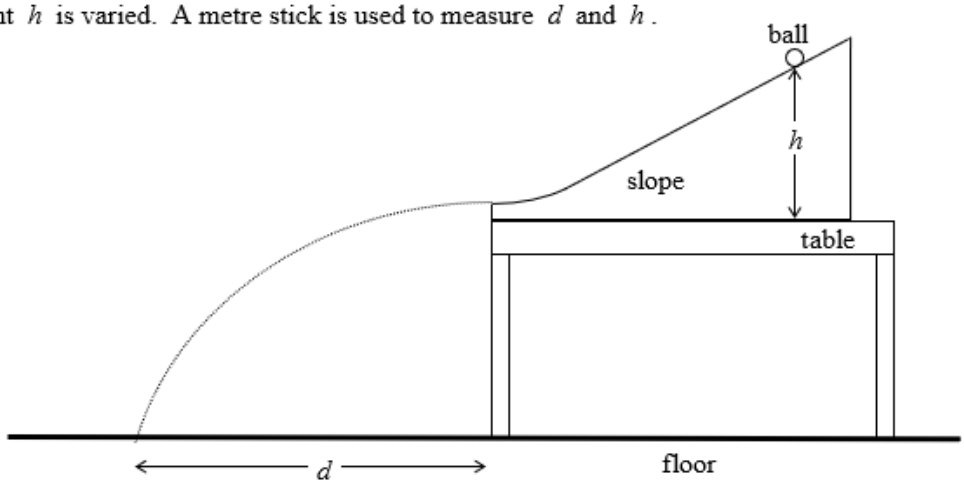
Write your student number in the box above.
This questionnaire is numbered up to page 6.
Read the text below and answer the questions on each sheet.
If you need more space for your answers, then use the backs of the sheets.

Please note that the data will be analysed anonymously and that you will not be identified to the lecturer of the course. The reason for asking you to fill in your student number is so that the person analyzing the response can contact you in order to clarify or expand on your responses.

Context:

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

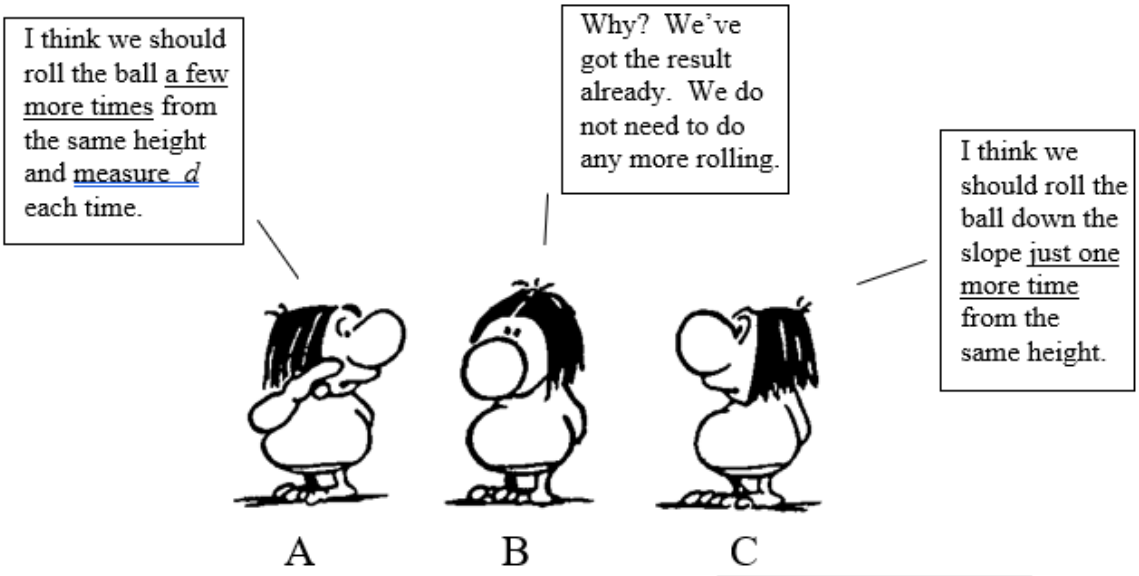
The students have been asked to investigate how the distance d on the floor changes when the height h is varied. A metre stick is used to measure d and h.



Q 1.

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

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A	B	C

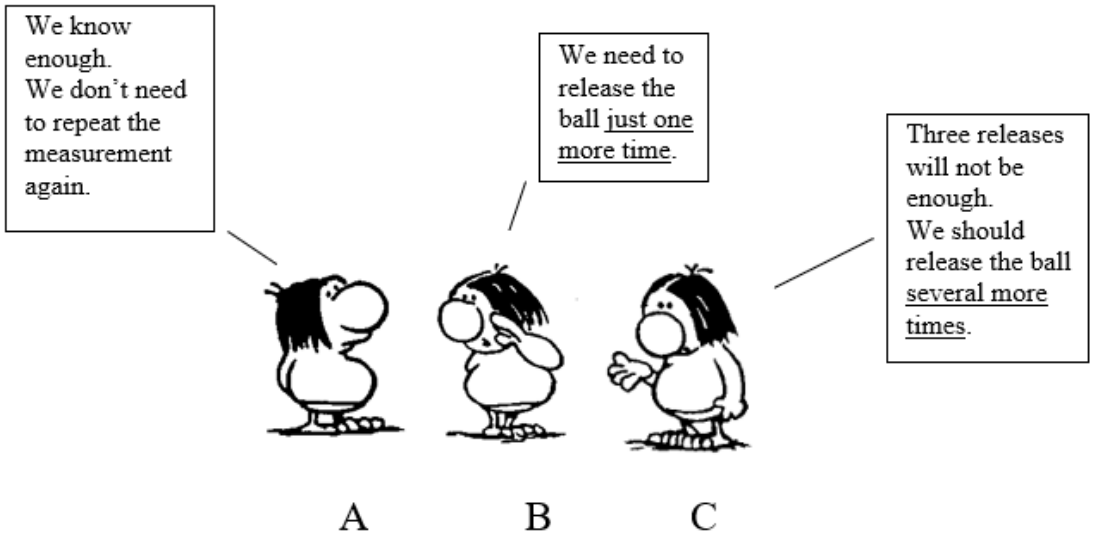
Explain your choice.

Q 2.

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

First release:	$h = 400$ mm	$d = 436$ mm
Second release:	$h = 400$ mm	$d = 426$ mm

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A	B	C

Explain your choice.

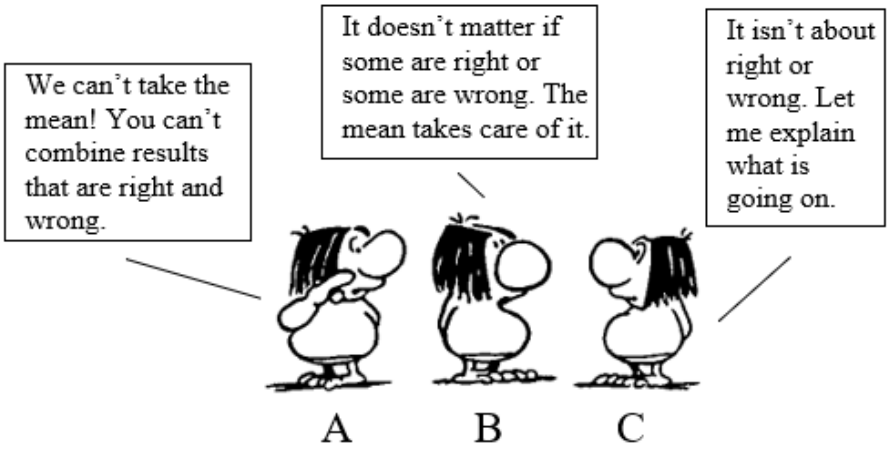
Q 3.

The students continue to release the ball down the slope at a height $h = 400$ mm.
They obtain the following after six release:

Release	d (mm)
1	436
2	425
3	440
4	425
5	434
6	425

One of the students says, "Great. We should now calculate the mean as the final result."

The following discussion then takes place between the students.



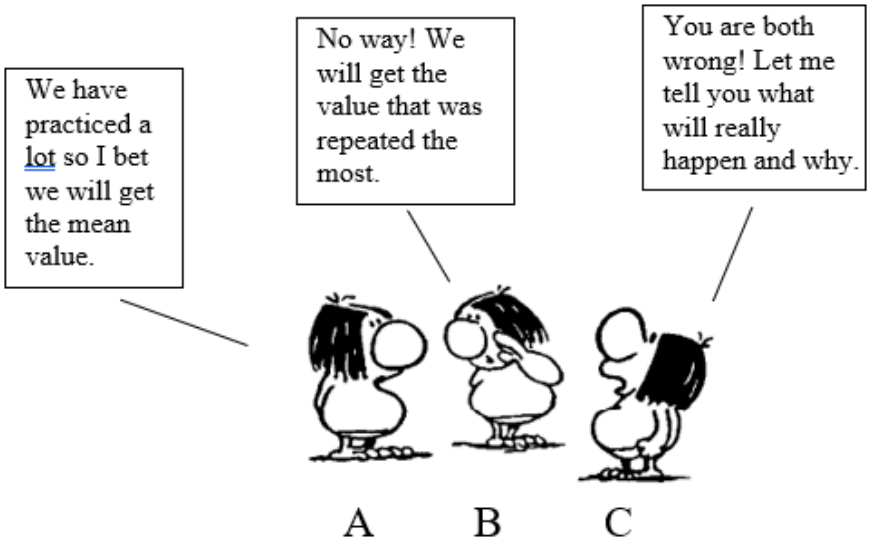
With whom do you most closely agree? (Circle ONE): | A | B | C |

Explain your choice.

Q 4.

The students take a bet on what result they will get for *d* if they release the ball again at a height $h = 400$ mm.

The following discussion then takes place between the students.

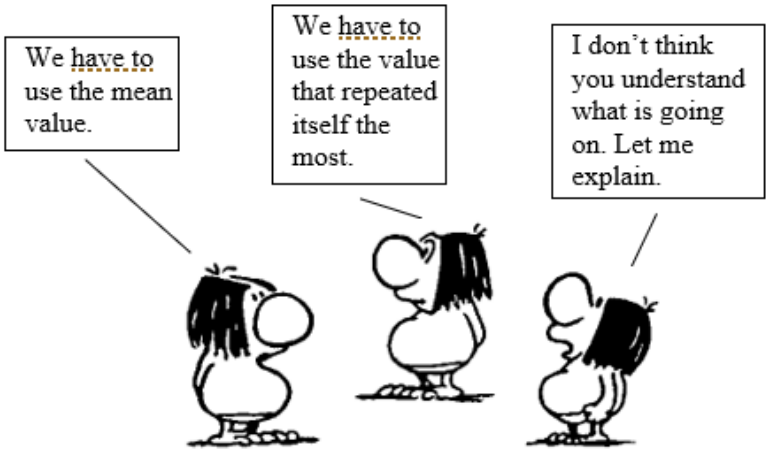


With whom do you most closely agree? (Circle ONE): | A | B | C |

Explain your choice.

Q 5.

The students then discuss what value to use for *d* in an equation.



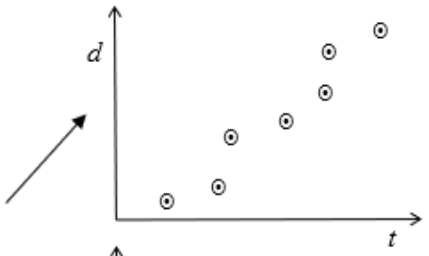
With whom do you most closely agree? (Circle ONE):

A	B	C

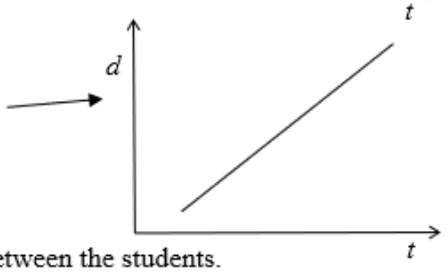
Explain your choice.

Q 6.

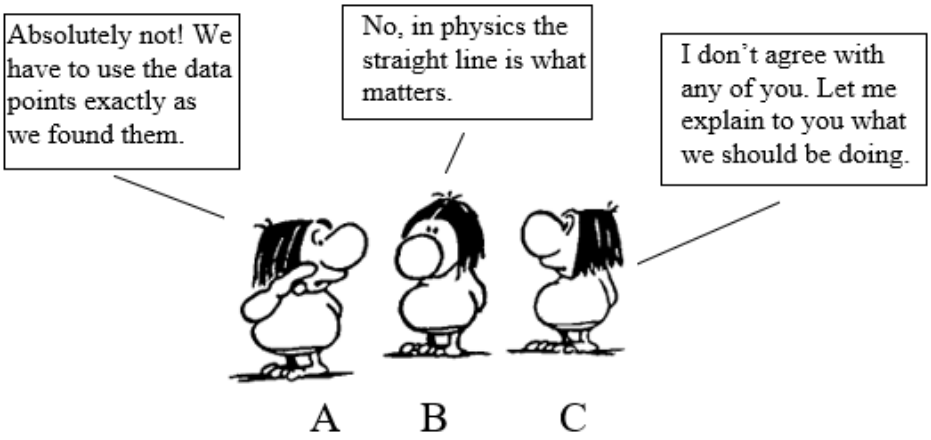
For the next part of the experiment they measure *how the time to hit the ground changes if the* the ball is released from *different heights*. They release the ball from 7 different heights *(h1, h2, h3... h7)* and measure the 7 corresponding times *(t1, t2 , t3...t7)*. They plot their data on a graph of *h* vs *t* as shown.



One of the students says, "I drew a line through the data. We must forget about the points now and use this!"



The following discussion then takes place between the students.



With whom do you <u>most closely</u> agree? (Circle ONE):

A	B	C

Explain your choice.

Appendix 2: Study 2's Instrument: S2

EMPLID:

S2

University of Cape Town
Department of Physics

Laboratory Questionnaire

Instructions:

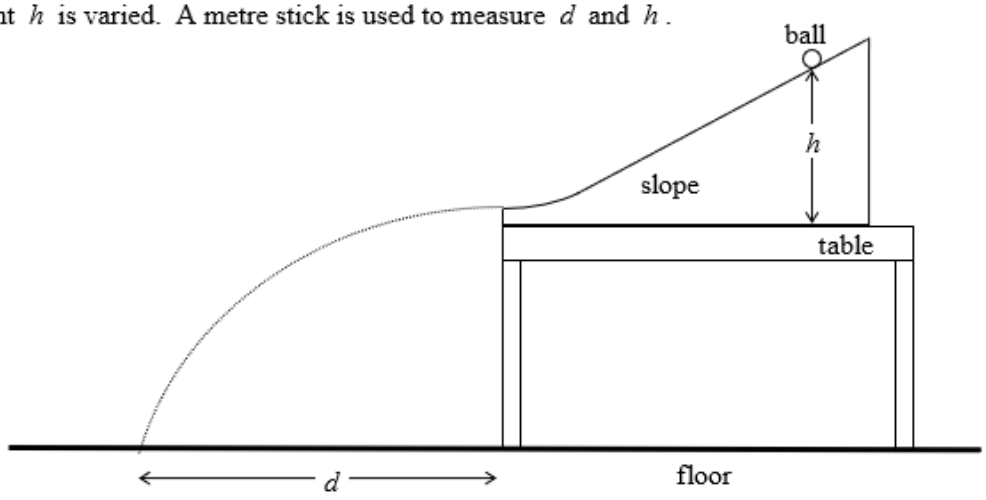
Write your Emplid in the box above.
This questionnaire is numbered up to page 4.
Read the text below and answer the questions on each sheet.
If you need more space for your answers, then use the backs of the sheets.

Please note that the data will be analyzed anonymously and that you will not be identified to the lecturer of the course. The reason for asking you to fill in your student number is so that the person analyzing the response can contact you in order to clarify or expand on your responses.

Context:

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

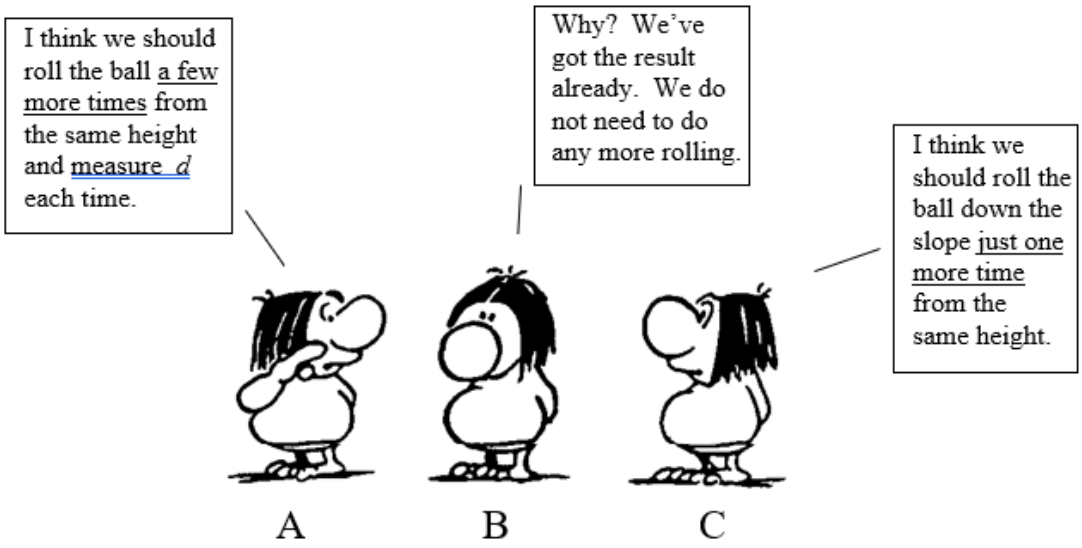
The students have been asked to investigate how the distance d on the floor changes when the height h is varied. A metre stick is used to measure d and h.



Q 1.

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

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A	B	C

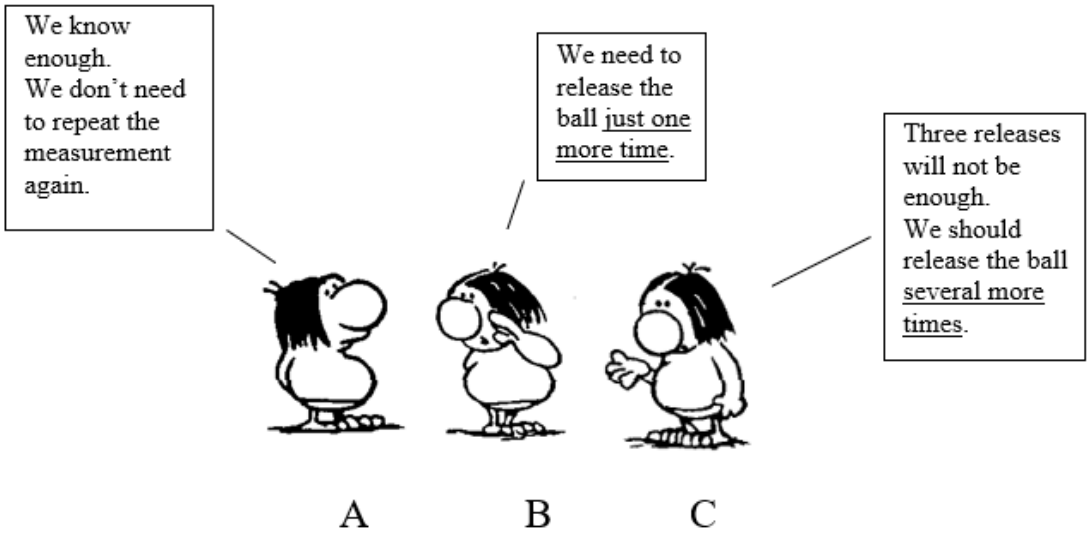
Explain your choice.

Q 2.

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

First release:	$h = 400$ mm	$d = 436$ mm
Second release:	$h = 400$ mm	$d = 426$ mm

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A	B	C

Explain your choice.

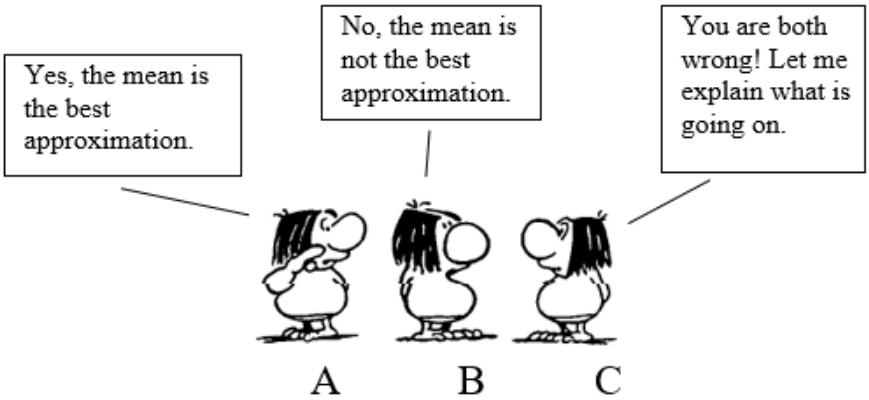
Q 3.

The students continue to release the ball down the slope at a height h = 400 mm.
They obtain the following after five releases:

Release	d (mm)
1	436
2	425
3	440
4	425
5	434

One of the students says, "Great. We can now calculate the mean as the final result."

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A → Explain carefully to your friend what you mean by "the mean is the best approximation".

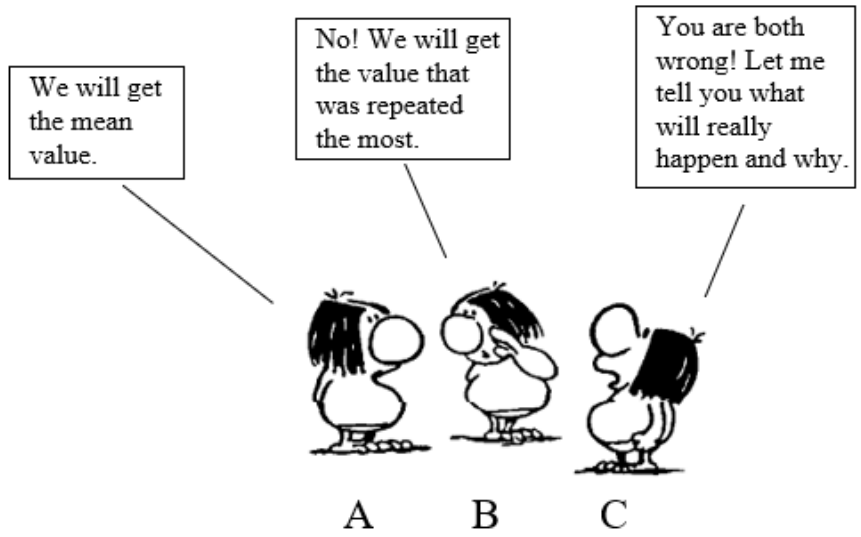
B → Explain carefully to your friend why you have this view.

C → Explain carefully to your friend why you have this view.

Q 4.

The students predict what they will get for d if they release the ball again at a height h = 400 mm.

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A	B	C

Explain your choice.

Appendix 3: Physics Measurement Questionnaire (PMQ)

SURNAME:	FIRST NAME:	
D/ [E]	University of Cape Town Department of Physics	Unique questionnaire number stamped here

Laboratory Procedures Questionnaire

Instructions:

Write your name in the box above.
Inside this envelope there are pages numbered up to page 10.
Read the text below and answer the questions on each sheet.
If you need more space for your answers, then use the backs of the sheets.
It should take you between 5 and 10 minutes to answer each question.

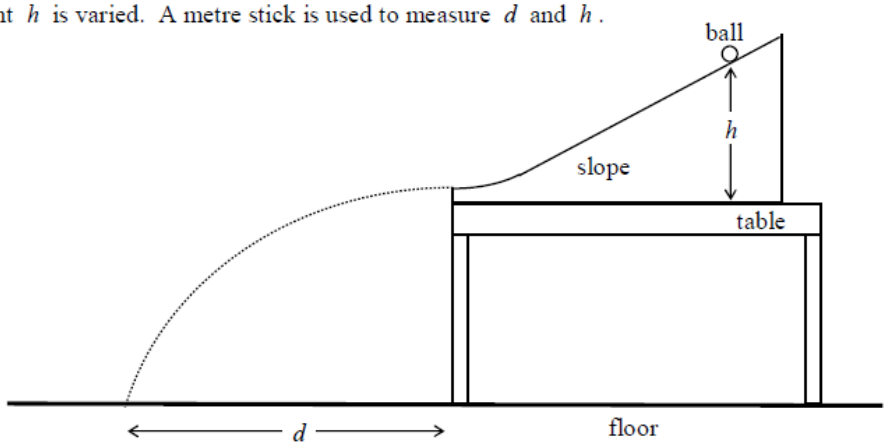
Answer the questions in order and do not skip any sheet.
When you have completed a question, put the sheet inside this envelope and do not take it out again, even if you want to change your answer.

Note: It is possible that some answers may be similar or exactly the same as others.
Please write all answers out in full, even if you feel that you are repeating yourself.

Context:

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

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

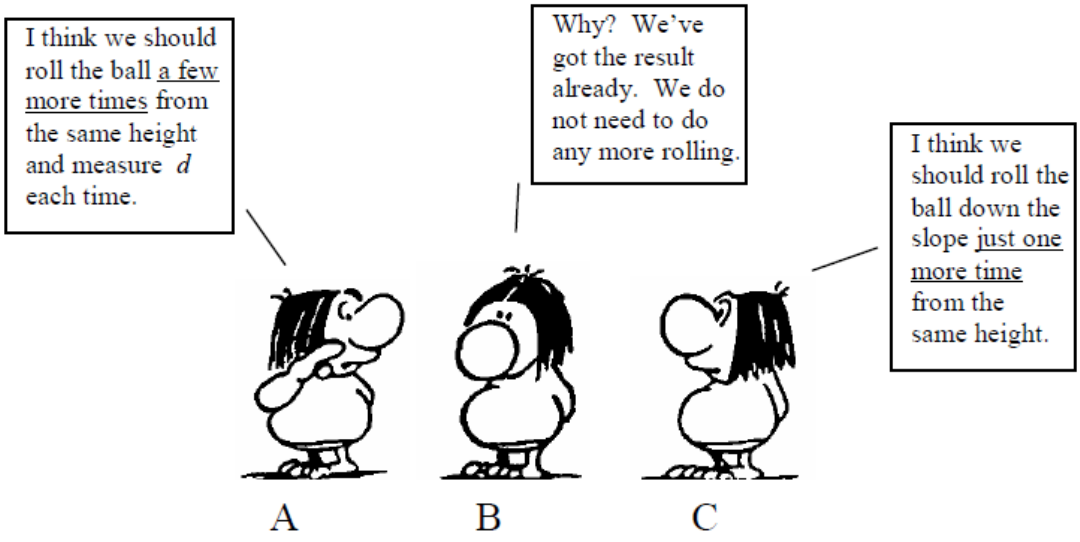


Unique questionnaire number stamped here

Q 1. (RD/D) [Q1. (RD/E) in post questionnaire]

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

The following discussion then takes place between the students.



With whom do you most closely agree? (Circle ONE):

A	B	C

Explain your choice.

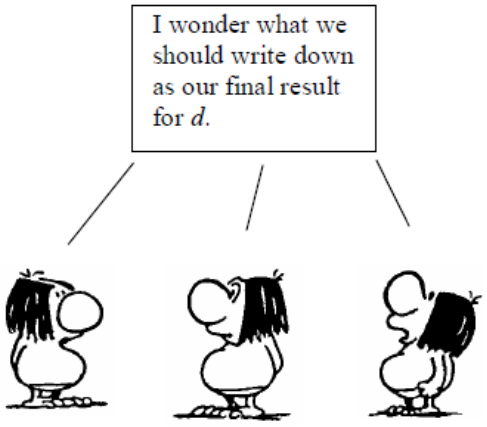
Unique questionnaire number stamped here

Q 3. (UR/D) [Q 4. (UR/E) in post questionnaire]

The students continue to release the ball down the slope at a height $h = 400$ mm.
Their results after five releases are:

Release	d (mm)	
1	436	
2	425	[426 in post questionnaire]
3	440	[438]
4	425	[426]
5	434	

The students then discuss what to write down for d as their final result.



Write down what you think the students should record as their final result for d.

Explain your choice.

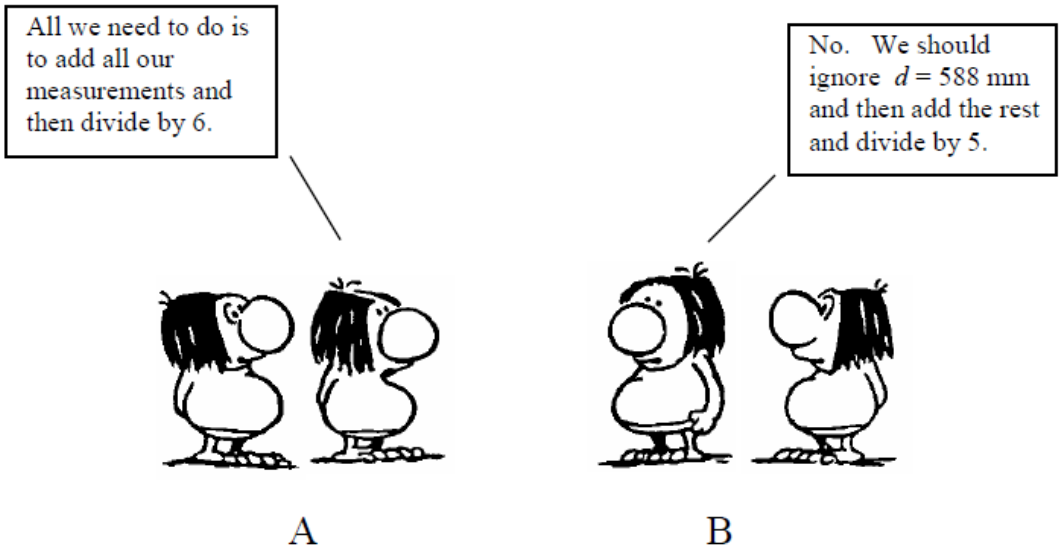
Unique questionnaire number stamped here

Q 4. (AN/D) [Q 5. (AN/E) in post questionnaire]

Another group of students have decided to calculate the average of all their measurements of d for $h = 400$ mm. Their results after six releases are:

Release	d (mm)
1	443
2	422
3	436
4	588
5	437
6	429

The students then discuss what to write down for the average of d.



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

Explain your choice.

_____

_____

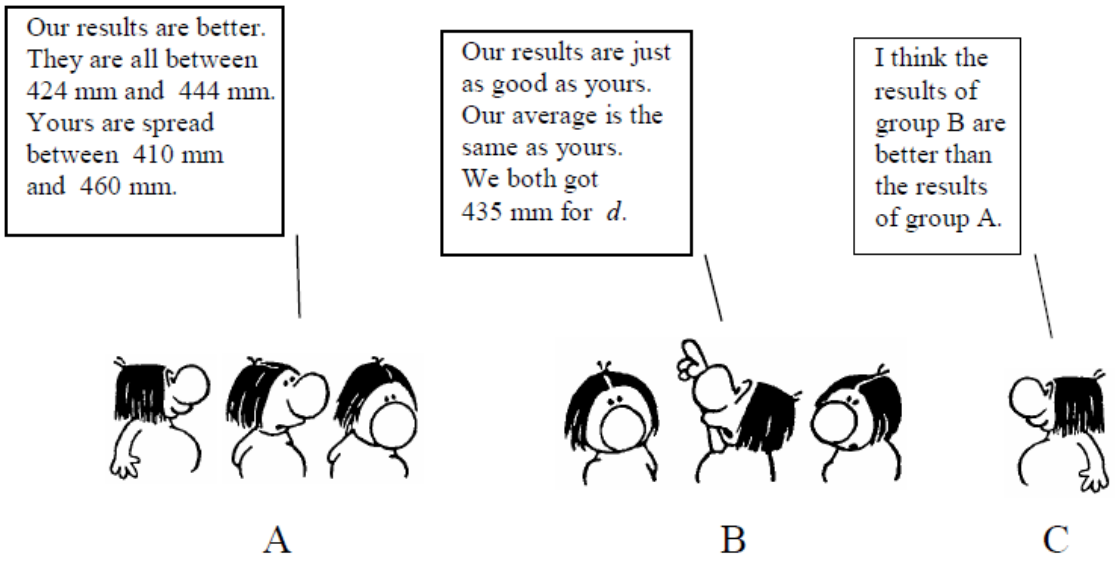
_____

Unique questionnaire number stamped here

Q 5. (SMDS/D) [Q 6. (SMDS/E) in post questionnaire]

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

Release	Group A d (mm)	Group B d (mm)
1	444	441
2	432	460
3	424	410
4	440	424
5	435	440
Average:	**435**	**435**



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

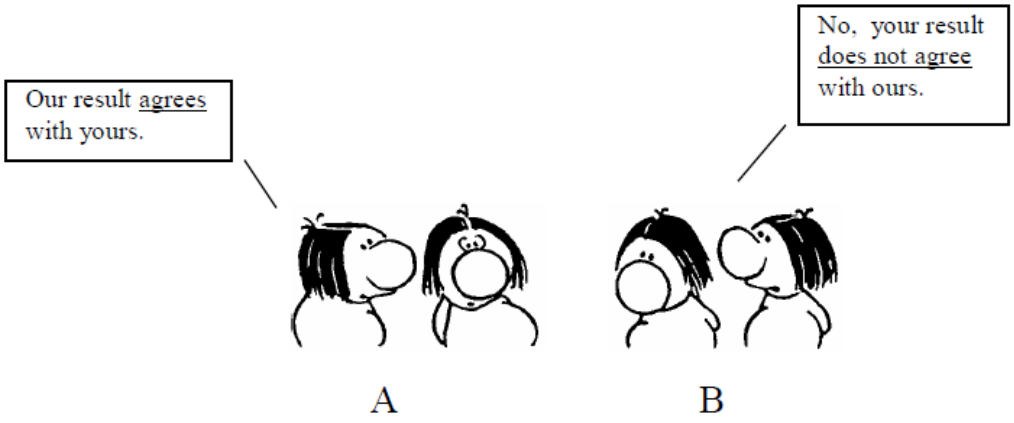
Explain your choice.

Unique questionnaire number stamped here

Q 6. (DMSS/D) [Q 7. (DMSS/E) in post questionnaire]

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

Release	Group A d (mm)	Group B d (mm)
1	440	432
2	438	444
3	433	426
4	422	433
5	432	440
Average:	**433**	**435**



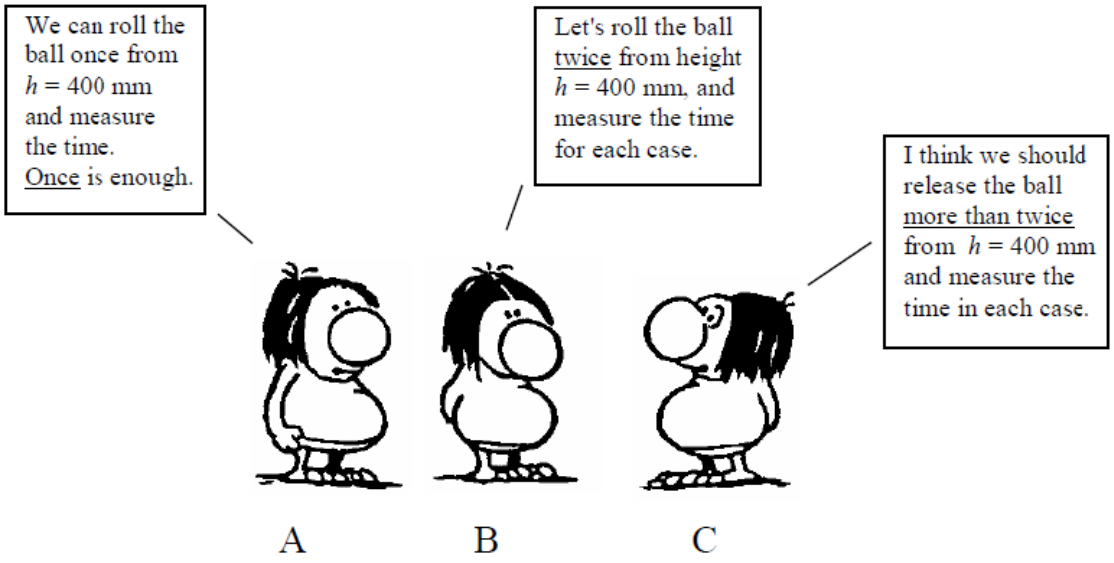
With which group do you most closely agree? (Circle ONE): A B

Explain your choice.

Unique questionnaire number stamped here

Q 7. (RT/D) { not used in post questionnaire }

The students are now given a stopwatch and are asked to measure the time that the ball takes from the edge of the table to hitting the ground after being released at $h = 400$ mm.
They discuss what to do.



With whom do you most closely agree? (Circle ONE):

A	B	C

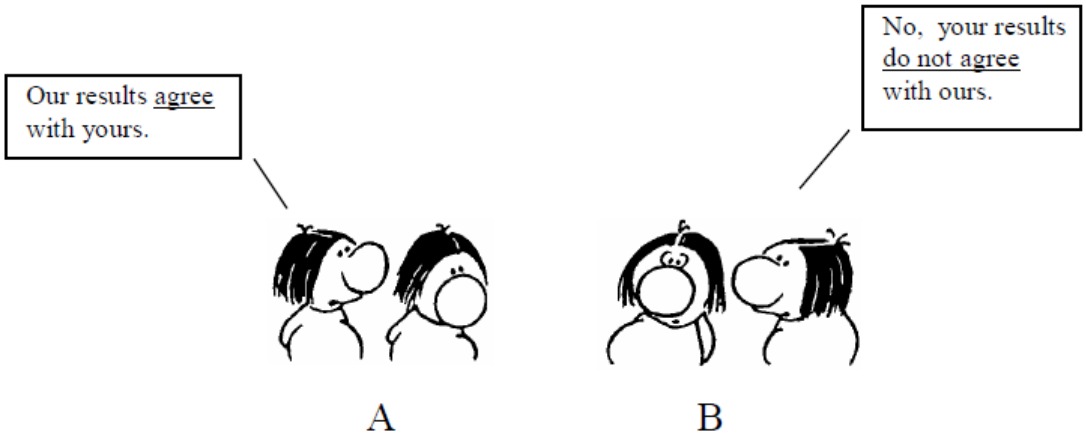
Explain your choice.

Unique questionnaire number stamped here

Q 8. (DMOS/E) { only administered in *post* questionnaire }

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

Release	Group A d (mm)	Group B d (mm)
1	444	458
2	435	438
3	424	462
4	440	449
5	432	443
Average:	**435**	**450**



With which group do you most closely agree? (Circle ONE): A B

Explain your choice. Do not use the word "results" in your explanation.

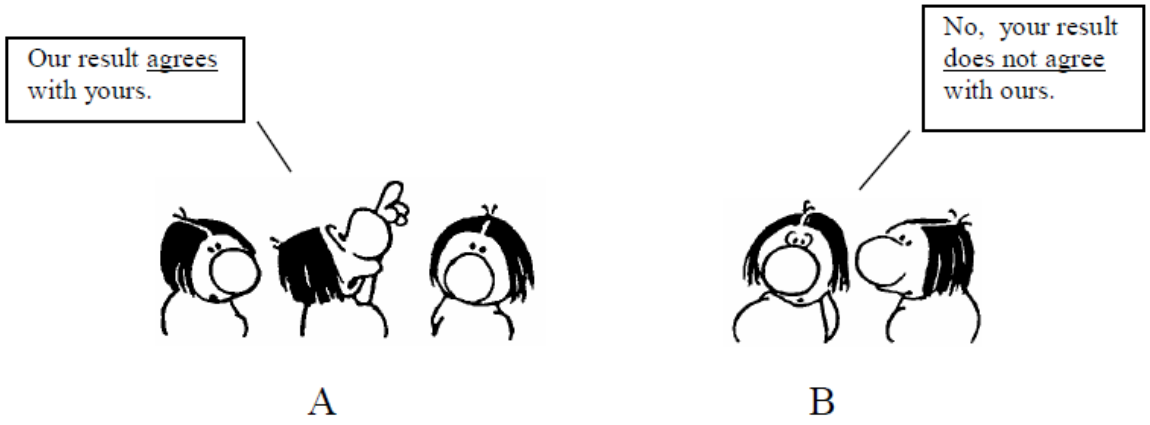
Unique questionnaire number stamped here

Q 9. (DMSU/E) { only administered in *post* questionnaire }

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

Group A: $d = 436 \pm 5$ mm

Group B: $d = 442 \pm 5$ mm



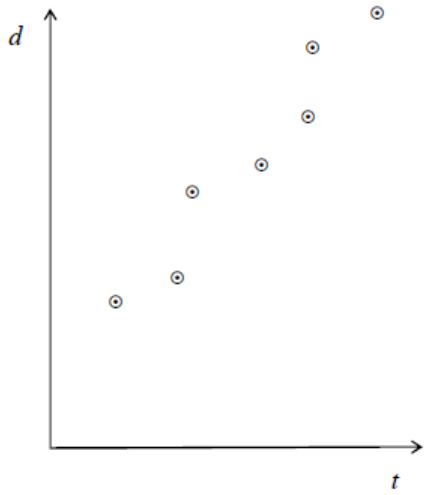
With which group do you most closely agree? (Circle ONE): | A | B |

Explain your choice. Do not use the word "result" in your explanation.

Unique questionnaire number stamped here

Q 8. (SLG/D) [Q 14. (SLG/E) in post questionnaire]

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



Explain carefully what you have done and why.

Unique questionnaire number stamped here

Q 9. [Q 15. in post questionnaire]

Comments.

Are there any answers to the previous question sheets that you want to change?
Please do not remove any sheets from the envelope.
What was the question about and how do you want to change your answer?

Any other comments?



In this laboratory questionnaire, I thought that the cartoon figures were (tick one):	male	female	mixed gender

Unique questionnaire number stamped here

Finally, please fill in these details: {this page not included in post questionnaire]

Surname:	**First names:**
Age:	Circle one: **Male** **Female**
Home language:	**Second language:**
Matric province:	**Name of School:**

Tick the **subjects** that you did in **matric**. Enter HG or SG and your symbol.
If you did subjects that are not listed, write them in the spaces provided.

Subject	**Tick**	**HG / SG**	**Symbol**
English first language			
English second language			
Mathematics			
Physical Science			
Biology			
History			
Geography			
Afrikaans			

Which programme have you registered on:

Student number:
(If you know it)

References

1. Allie, S., Demaree, D., Taylor, J., Lubben, F., & Buffler, A. (July, 2008). Making sense of measurement, making sense of the textbook. Paper presented at the *Physics Education Research Conference,* 3-6.

2. Allie, S., & Buffler, A. (1998). A course in tools and procedures for Physics I. *American Journal of Physics, 66*(7), 613-624. doi:10.1119/1.18915

3. Allie, S., Buffler, A., 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. doi:10.1080/0950069980200405

4. Allie, S., Buffler, A., Campbell, B., & Lubben, F. (1999). Procedural understanding of pre-first year science students at the University Of Cape Town, South Africa. *Proceedings of the 1998 6th International Symposium of The Oxford Centre for Staff and Learning Development,* 146-156.

5. Allie, S., Buffler, A., Campbell, B., Lubben, F., Evangelinos, D., Psillos, D., & Valassiades, O. (2003). Teaching measurement in the introductory physics laboratory. *The Physics Teacher, 41*(7), 394-401. doi:10.1119/1.1616479

6. Allie, S., Buffler, A., Kaunda, L., & Inglis, M. (1997). Writing-intensive physics laboratory reports: Tasks and assessment. *The Physics Teacher, 35*(7), 399-405. doi:10.1119/1.2344739

7. Allie, S., Buffler, A., Lubben, F., & Campbell, B. (2002). Point and set paradigms in students' handling of experimental measurements. *Science Education: Past, Present and Future,* doi:10.1007/0-306-47639-8_47

8. Allie, S., & Demaree, D. (2010). Toward meaning and scientific thinking in the traditional freshman laboratory: Opening the “Idea space”. Paper presented at the *Physics Education Research Conference,* 1-4.

9. Bok, J. (2014). Probing student views about the big bang in an introductory astronomy course. *Unpublished honours project: University of Cape Town*.

10. Buffler, A., Allie, S., Campbell, B., & Lubben, F. (1998). The role of laboratory experience on the procedural understanding of pre-first year science students at UCT. *Proceedings at the Sixth Annual Meeting of the Southern African Association for Research in Mathematics and Science Education,* 469-502.

11. Buffler, A., Allie, S., & Lubben, F. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education, 23*(11), 1137-1156. doi:10.1080/09500690110039567

12. Buffler, A., Lubben, F., & Ibrahim, B. (2009). The relationship between students' views of the nature of science and their views of the nature of scientific measurement. *International Journal of Science Education, 31*(9), 1137-1156. doi:10.1080/09500690802189807

13. Campbell, B., Lubben, F., Buffler, A., & Allie, S. (2005). Teaching scientific measurement at university: Understanding student ideas and laboratory curriculum reform. *South African Association for Research in Mathematics, Science and Technology Education.*

14. Coelho, S. M., & Séré, M. (1998). Pupils' reasoning and practice during hands-on activities in the measurement phase. *Research in Science & Technological Education, 16*(1), 79-96. doi:10.1080/0263514980160107

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

16. Evangelinos, D., Psillos, D., & Valassiades, O. (2002). An investigation of teaching and learning about measurement data and their treatment in the introductory physics laboratory. *Science Education Research in the Knowledge Based Sciences, 1,* 380-382.

17. Fairbrother, R., & Hackling, M. (1997). Is this the right answer? *International Journal of Science Education, 19*(8), 887-894. doi:10.1080/0950069970190802

18. Gott, R., Duggan, S., Miller, R., & Lubben, F. (1995). Progression in investigative work in science. *Subject Learning in the Primary Curriculum: Issues in English, Science and Mathematics. London: Routledge*

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

20. Gott, R., & Duggan, S. (1996). Practical work: Its role in the understanding of evidence in science. *International Journal of Science Education, 18*(7), 806. doi:10.1080/0950069960180705

21. Hammer, D. (1994). Epistemological beliefs in introductory physics. *Cognition and Instruction, 12*(2), 151-183. doi:10.1207/s1532690xci1202_4

22. Hestenes, D. (1992). Modelling games in the Newtonian world. *American Journal of Physics, 60*(8), 732-748. doi:10.1119/1.17080

23. Ibrahim, B., Buffler, A., & Lubben, F. (2009). Profiles of freshman physics students' views on the nature of science. *Journal of Research in Science Teaching, 46*(3), 248-264. doi:10.1002/tea.20219

24. Kanari, Z., & Miller, R. (2003). How children reason from data to conclusions in practical science investigations. *Science Education Research in The Knowledge-Based Society, ,* 117-125.

25. Leavy, A., & O'Loughlin, N. (2006). Preservice teachers understanding of the mean: Moving beyond the arithmetic average. *Journal of Mathematics Teacher Education, 9*(1), 53-90. doi:10.1007/s10857-006-9003-y

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

27. Lubben, F., Campbell, B., Buffler, A., & Allie, S. (2001). Point and set reasoning in practical science measurement by entering university freshmen. *Science Education, 85*(4), 311-327. doi:10.1002/sce.1012

28. Majiet, N. (2016). Student views on the role of experiments in physics. *Unpublished honours project: University of Cape Town*.

29. Majiet, N., & Allie, S. (2018). Student understanding of measurement and uncertainty: Probing the mean. Paper presented at the *Physics Education Research Conference 2018*

30. Masnick, A. M., & Morris, B. J. (2002). Reasoning from data: The effect of sample size and variability on children's and adults' conclusions. *Proceedings of the Annual Conference of the Cognitive Science Society,* (24), 643-648.

31. Masnick, A. M., & Klahr, D. (2003). Error matters: An initial exploration of elementary school children's understanding of experimental error. *Journal of Cognition and Development, 4*(1), 67-98. doi:10.1080/15248372.2003.9669683

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

33. 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.

34. Pillay, S., Buffler, A., Lubben, F., & Allie, S. (2008). Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory. *European Journal of Physics, 29,* 647-659.

35. Rajpaul, V., Allie, S., & Blyth, S. (2014). Introductory astronomy course at the University of Cape Town: Probing student perspectives. *Physical Review Special Topics - Physics Education Research, 10*(2), 020126. doi:10.1103/PhysRevSTPER.10.020126

36. Séré, M., Journeaux, R., & Larcher, C. (1993). Learning the statistical analysis of measurement errors. *International Journal of Science Education, 15*(4), 427-438. doi:10.1080/0950069930150406

37. Tlowana, M. (2017). Student perceptions of the introductory physics laboratory: An exploratory study. *Unpublished master's thesis: University of Cape Town*.

38. Tsai, C. (1998). An analysis of scientific epistemological beliefs and learning orientations of Taiwanese eighth graders. *Science Education, 82*(4), 473-489. doi:AID-SCE4>3.0.CO;2-8

39. Vellom, R. P., & Anderson, C. W. (1999). Reasoning about data in middle school science. *Journal of Research in Science Teaching, 36*(2), 179-199. doi:10.1002/(SICI)1098-2736(199902)36:2<179::AID-TEA5>3.0.CO;2-T

40. Volkwyn, T. (2005). First year students' understanding of measurement in physics laboratory work. *Unpublished master's thesis: University of Cape Town.*

41. Volkwyn, T. S., Allie, S., Buffler, A., & Lubben, F. (2008). Impact of a conventional introductory laboratory course on the understanding of measurement. *Physical Review Special Topics - Physics Education Research, 4*(1), 010108. doi:10.1103/PhysRevSTPER.4.010108

42. Watson, J. (2018). Variation and expectation for six-year-olds. *Statistics in early childhood and primary education: Supporting early statistical and probabilistic thinking* (pp. 55-73). Singapore: Springer. doi:10.1007/978-981-13-1044-7

43. Watson, J. M., & Kelly, B. A. (2005). The winds are variable: Student intuitions about variation. *School Science and Mathematics, 105*(5), 252-269. doi:10.1111/j.1949-8594.2005.tb18165.x

44. Wilcox, B. R., & Lewandowski, H. J. (2018). A summary of research-based assessment of students' beliefs about the nature of experimental physics. *American Journal of Physics, 86*(3), 212-219. doi:10.1119/1.5009241