IOP PUBLISHING EUROPEAN JOURNAL OF PHYSICS

Eur. J. Phys. **29** (2008) 647–659 doi:10.1088/0143-0807/29/3/024

Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory

Seshini Pillay, Andy Buffler, Fred Lubben and Saalih Allie

Department of Physics, University of Cape Town, Rondebosch, 7701, South Africa

E-mail: andy.buffler@uct.ac.za

Received 19 December 2007, in final form 25 February 2008
Published 29 April 2008
Online at stacks.iop.org/EJP/29/647

Abstract

An evaluation of a course aimed at developing university students' understanding of the nature of scientific measurement and uncertainty is described. The course materials follow the framework for metrology as recommended in the *Guide to the Expression of Uncertainty in Measurement* (GUM). The evaluation of the course is based on responses to written questionnaires administered to a cohort of 76 first year physics students both pre- and post-instruction, which were interpreted in terms of 'point' or 'set' reasoning. These findings are compared with responses from a control group of 70 students who completed a similar laboratory course apart from the use of traditional approaches to measurement and data analysis. The results suggest that the GUM framework, together with the specific teaching strategies described, provides opportunities for more effective learning of measurement and uncertainty in the introductory laboratory.

1. Introduction

Although most undergraduate physics courses include a laboratory component, there is little consensus about the rationale behind the inclusion of this practical element [1–4]. Amongst the purposes that are put forward are the demonstration of physical principles introduced in lectures, the provision of 'hands-on' opportunities to familiarize students with experimental procedures and apparatus, and the practice of the 'scientific method' including communication through scientific report writing [5, 6]. One common thread that does, however, run through most introductory laboratory courses is that experimentation involves taking and analysing measurements.

Although such courses often include an explicit presentation of some of the key elements involved in carrying out the data analysis, there is an implicit assumption that by repeating a

number of well-defined algorithms, students will also come to acquire a deeper understanding of the nature of scientific measurement and uncertainty. This assumption has, however, been challenged in a variety of contexts [7–12]. For example, it has been shown [7–9] that although the majority of students may end up being able to correctly apply procedures such as calculating means and standard deviations, the same students may in fact decide to repeat measurements only if they have reasons to query their first reading, and then with the purpose of finding a recurring value. In another study [10] it was found that students included anomalous readings in calculations of the mean without any comment, even when the reading differed from the next closest reading by more than 8 standard deviations. Students also neglected to interpret and use their calculated uncertainties and hence compared results arbitrarily [10].

Apropos of these concerns, the effectiveness of some new teaching approaches that aim to develop students' understanding of measurement in laboratory courses has been reported [9, 11–13]. While several of these [9, 11, 12] showed significant increases in students' understanding of measurement in data collection and data processing, the improvement in students' understanding of uncertainty, particularly when comparing sets of results, was less satisfactory. Thus, while students may execute the technical procedures of data analysis adequately, they often appear to lack the conceptual basis from which the computational aspects are derived. One of the reasons suggested for this [14] is that the difficulties students experience are directly related to difficulties that are inherent in the approaches traditionally used to teach data analysis. In this paper, we report on an evaluation of an innovative course on measurement which was designed to improve students' understanding of measurement and uncertainty within the context of the introductory physics laboratory.

2. Conceptual difficulties in the traditional approach to teaching data analysis

The traditional approach to data analysis, which is used or implied in most undergraduate laboratories, relies on analysing data in terms of frequencies, and is hence often termed 'frequentist'. For meaningful data analysis to be carried out within this formalism, a large enough data set, of the order of 20 observations or more, is required [15]. Thus, the formalism inherently offers no prescription for modelling a single datum. In addition, traditional instruction usually emphasizes 'random error' for which there is a rigorous mathematical procedure, while 'systematic errors' are simply treated as unknown constants that are determined by examining the experimental set-up. The result is that a number of *ad hoc* prescriptions and rules of thumb are introduced that militate against students putting together a meaningful and coherent framework.

From a teaching point of view it has become common to foreground the notion of significant figures, and in many instances this idea is used to introduce the concept of uncertainty. Thus, a purely descriptive term which should arise as a consequence of having calculated an uncertainty interval is often elevated to the status of a primary argument. It is also unfortunate that the terminology that is used in measurement often conjures up incorrect associations in the minds of the novice [16]. For example the term 'error' misleads students by suggesting the existence of true and false experimental results, possibly reinforcing the idea that an experiment has one predetermined 'correct' result known by the instructor [10], while students' measurements are often 'in error'. The phrase 'due to human error' often appears in students' laboratory reports even after significant exposure to laboratory work. Similarly, the use of the term 'uncertainty' has the everyday connotations of not being sure, thus potentially inducing a technically incorrect perspective about the nature of measurement. These issues have made both the teaching and learning of experimentation difficult.

Table 1. Outline of the content of the interactive student workbook.

Unit	Description
1. Introduction to measurement	The relationship between science and experiment. The nature and purpose of measurement.
2. Basic concepts of measurement	Probability and inference. Reading digital and analogue scales. The nature of uncertainty. A probabilistic model of measurement.
3. The single reading	Probability density functions. Representing knowledge graphically using a pdf. Evaluating standard uncertainties for a single reading. The result of a measurement.
4. Repeated readings that are dispersed	Dispersion in data sets. Evaluating standard uncertainties for multiple readings. Type A and Type B evaluation of uncertainties.
5. Working with uncertainties	Propagation of uncertainties. Combined standard uncertainty. The uncertainty budget. Comparing different results. Repeatability and reproducibility.
6. Modelling trends in data	Principle of least squares. Least-squares fitting of straight lines.

3. Our probabilistic course for measurement and uncertainty based on the GUM framework

The fragmented way of applying the frequentist formalism and terminology across different science disciplines led the *Bureau International des Poids et Mesures* to review the situation with regard to calculating and reporting measurements and uncertainties [17]. These efforts culminated in the issue of the *Guide to the Expression of Uncertainty in Measurement* (GUM) [18, 19] by the *International Organization for Standardization* (ISO), which has now been adopted by all international standards organizations including the *International Union of Pure and Applied Physics* and the *National Institute of Standards and Technology* [20]. One of the key features of the ISO-advocated probabilistic approach is that measurement is viewed as a problem of inference, and probability theory [21] is used to construct claims about a measurand based on the information at hand. Soon after the publication of the ISO recommendations in 1995 it was suggested that this probabilistic approach was not only required for scientific enquiry in experimental work but should also guide the teaching of the techniques for data treatment [14].

We have designed and written a set of materials in the form of a student workbook [22] based on the framework of measurement and uncertainty as specified by the GUM. The materials are also informed by our research findings that relate to students' understanding of measurement [8, 9]. The broad content areas in the workbook are listed in table 1. The concept of the 'measurand' is introduced early on together with exercises that deal explicitly with the difference between a reading that is observed on a measuring instrument and the value of the measurand that can then be inferred. One of the recommendations in the GUM [18] that lends itself to the teaching situation is that an 'uncertainty budget' should be part of all reported experiments. This idea is introduced qualitatively at first by asking students to reflect on the experiments they are doing, by writing down all the factors that could have influenced their results and then to judge whether they think these influences would have a 'large' or 'small' effect. Measurements based on a single observation are introduced before dealing with ensembles of data with scatter, since dispersion may then be introduced as one of the many sources of uncertainty, and not necessarily the dominant one. The idea that the probability density function (pdf) is a tool which represents all the available information about a particular measurand is illustrated in a variety of contexts. The best approximation

and the standard uncertainty are introduced as the two parameters which can be used to summarize the information associated with a particular pdf. The final stage in the teaching sequence deals with formulating the result of a measurand in terms of a probabilistic statement. The examples in the workbook guide students through a range of measurement contexts, calculating standard uncertainties for a variety of sources of uncertainty and 'summing' these to provide a combined standard uncertainty for the measurement. In this way, the theme of considering and evaluating every possible source of uncertainty culminates in the students being able to draw up uncertainty budgets and to determine a reasonable total uncertainty for their measurements in practical tasks.

The measurement course [22] was piloted in the Physics Department at the University of Cape Town in 2002 and has been running since in a variety of our first year physics courses. In our context, the laboratory programme consists of one 3 h session per week for 24 weeks spanning the entire teaching year, with about eight of these sessions being devoted to the measurement course. Experience shows that tutors need to be trained with respect to the use of the probabilistic approach in order to prevent them resorting to the (familiar) frequentist approach for data treatment, which underpinned their own laboratory training. Also, the practical tasks used in the laboratory course should be designed purposely to allow for issues of measurement and uncertainty to emerge, and the instructions for the practical tasks should reflect the GUM notation and approaches in a consistent way, particularly the use of uncertainty budgets. In our course, laboratory tasks have been framed in the form of problems that require experimental investigation for their resolution, thus providing a meaningful purpose for measurement and data analysis.

The remainder of this paper reports on the evaluation of the course guided by the following questions:

(i) to what extent does the course based on the GUM framework improve students' understanding of scientific measurement and uncertainty and
(ii) how effective is the new course in this regard compared to a course incorporating traditional (frequentist) approaches to data analysis?

4. Evaluation of the course: methodology

The evaluative data came from responses of 76 students to written questionnaires administered before and after their laboratory course. These students followed a structured 4 year BSc programme primarily targeted at educationally disadvantaged black students in South Africa. Typically these students have had little or no laboratory experience at high school, do not speak English as a first language and tend to come from socio-economically depressed backgrounds. A previous cohort of 70 students was used as a control group. The experimental and control groups had similar demographic characteristics, entered for the same physics programme, and experienced similar practical tasks in their laboratory courses. However, the experimental group was exposed to the new course on measurement based on the GUM, while the control group used traditional frequentist procedures for data treatment.

The pre-instruction questionnaire [23] was composed of eleven items, or 'probes', nine of which will be reported on in this study, and the post-instruction questionnaire [23] comprised nine similar probes. Each questionnaire was based on one simple experimental scenario, and each individual probe provided a brief description of a practical measurement situation arising in the particular scenario and several possible courses of action. The student was required to make a choice and then provide a justification for the choice. The probes had been independently validated [24] or previously piloted for appropriateness of language level and

Table 2. Descriptors defining the point and set paradigms.

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

diversity of responses. Each probe was answered individually in strict sequence and under examination conditions [8] typically taking a total of 45 min to complete the questionnaire. A large-scale version of the apparatus was also used to demonstrate the experiment before the probes were answered. Each probe targets a particular aspect of measurement: three probes deal with a single reading, two probes focus on processing sets of data, another two probes deal with the comparison of two independent sets of data and a final probe explores views on the nature of uncertainty.

Responses were classified according to whether or not the declared idea was compatible with the point paradigm or the set paradigm [9]. Table 2 lists the key descriptors for the two paradigms. The point paradigm is characterized by the notion that each experimental observation potentially yields the correct or 'true' value of the quantity being measured (the measurand). Consequently, each observation is considered independent of every other observation. A single carefully performed observation is considered sufficient to establish the true value of the measurand. Any deviation from an expected result is attributed to environmental influence or mistakes made by the experimenter. Where an ensemble of readings does exist (with dispersion in the values), the measurement result is obtained by selecting one of the data points. In the cases where independent data sets need to be compared for agreement or the lack thereof, a one-to-one comparison of individual observations is used.

The set paradigm, on the other hand, is characterized by the notion that each observation provides partial information about the measurand. Thus, all available data are used to construct a distribution from which the best approximation of the measurand and an interval of uncertainty are derived. Most commonly, the best approximation of the measurand will be the observation itself (for a single reading), or the average of the readings (for an ensemble of readings with dispersion). Comparisons between different sets of results involve comparing to what extent the intervals overlap or not.

The increase in the proportion of students using ideas associated with the set paradigm was taken as a criterion for the effectiveness of the new course based on the GUM framework of measurement.

The first stage of the analysis of the questionnaires involved two researchers independently developing categories of responses [23] for each probe, thus increasing coding scheme validity. Each response category was assigned an alphanumeric code based on the student's choice of action and supporting explanation. Again, two researchers independently used the coding scheme for allocating codes to half of the responses for each probe, with an inter-coder agreement of 84%. The frequencies of responses for each probe were tallied and clusters of responses showing similar types of reasoning were identified. Each category of responses was associated with either the point or set paradigm. This enabled the underlying reasoning to be identified for each student across different measurement-related situations.

For the students in the experimental sample, frequencies of the use of both paradigms before and after the new course were compared for their ideas of measurement in four contexts, i.e. when collecting data through a single observation, processing multiple observations, comparing data sets and expressing the nature of uncertainty. For students in the experimental and control groups the consistency of the use of both paradigms at the end of the courses was compared in order to establish the overall effectiveness of the new GUM-compliant course.

5. Evaluation of the course: findings

5.1. Students' ideas about data collection through multiple observations

A number of previous studies [8–11] have shown that regardless of the nature of the laboratory course, nearly all students believe it necessary to undertake repeated observations of the same phenomenon, if possible. We have therefore excluded the two probes testing these views from the evaluation of this course. Students' pre-instruction reasoning for these probes is nevertheless interesting and has been reported elsewhere [8, 25].

5.2. Students' ideas about data collection through a single observation

Three probes were used to explore students' views on the relationship between a single observation and the measurand. The first probe presented a ruler with a reference point apparently positioned directly below a graduation mark. The second probe also presented a ruler, but with the reference point between two graduation marks. The third probe required the interpretation of a single reading of a digital scale. Responses from both the pre- and post-intervention questionnaires are used below to illustrate reasoning associated with the point and set paradigms.

The majority of responses associated with the point paradigm indicated that the true value is attainable from a measuring device with a finely calibrated scale, as illustrated by the following responses:

> The distance is exactly 436 mm as there are 10 subdivisions in between 430 and 440 and the mark is on the 6th one.

> The distance d is approximately 426.5 mm, because the ruler is not adequately measured in units less than mm. So it is sufficient to say we can't know for sure if it did indeed travelled 426.5 mm because we don't know for sure if it is a 0.5 mm.

Point reasoning evident when reporting a digital reading was typically manifested in a claim that the digital scale shows the exact (true) value, as illustrated by the response:

> The distance is exactly 423.7 mm, since electronic things always make no mistakes.

On the other hand, responses associated with the set paradigm typically alluded to the incomplete information obtained from a single analogue or digital reading, as demonstrated by this response:

> In Physics we can never get an exact distance, only an approximate distance close to the real one.

Other responses indicated that variation in the readings is due to the experimental set-up and calibration, as suggested in the interpretation of a digital reading below:

> You can't really say what the distance is. Even an electronic meter cannot say accurately what in each case the result will be, it only sorts out the case so you are accurate when measuring the distance the ball landed each time, but external factors, and the position the ball was released could still affect that distance d.

Table 3. Students' reasoning when dealing with a single reading ($n = 76$).

	Pre-instruction reasoning (%)			Post-instruction reasoning (%)		
Probe description	Point	Set	Neither	Point	Set	Neither
Reading a ruler with mark on a scale graduation	45 (59)	17 (23)	14 (18)	21 (27)	55 (74)	0 (0)
Reading a ruler with mark between two scale graduations	44 (57)	23 (31)	9 (12)	23 (31)	52 (69)	1 (1)
Reading a digital scale	39 (52)	31 (40)	6 (8)	7 (9)	67 (84)	2 (3)

Several post-intervention responses explicitly emphasized the need to calculate the best estimate and standard uncertainty, as indicated by this inference of a digital reading:

> The value is between 432.5 and 433.5 mm, because when using a digital instrument we can say that the uncertainty of a value is $u(d) = 0.29$ and therefore $d = 433.0 \pm 0.29$ which does fall within 432.5 and 433.5. We have defined an interval in which the true value lies as 433.0 is our best estimate.

Table 3 summarizes students' reasoning when dealing with single readings before and after instruction. Before instruction more than half the students used reasoning associated with the point paradigm for all three probes, and after instruction this proportion decreased to 27% and 31% for the two analogue readings, respectively, and 9% for the digital reading. In contrast, reasoning associated with the set paradigm more than doubled for each probe after instruction. A considerable proportion of the pre-instruction responses were not classifiable, mainly due to the students' ambiguous use of terms such as 'accurate' and 'results'.

5.3. Students' ideas about processing multiple observations

Three probes were used to elicit students' views on the relationship between a set of repeated observations and the measurand. One probe presented a set of readings (including two identical ones) and asked for the best number to represent the data set. A second probe presented a set of readings with its average value and asked students to interpret this information. The last probe in this cluster asked for a prediction of a subsequent reading given a set of five readings with its average value.

Point reasoning was evident when the student was asked to represent a set of repeated readings and expressed the idea that the best estimate of the measurand was the recurring value in the data set:

> I would say 434 mm because it has landed there more than once. Meaning it hit that spot at the same speed at the same times, so it has to be the right distance.

Point reasoning was apparent in the other probes when the average was seen to represent the exact result, as illustrated by these two quotes:

> I think the distance is exactly 432 mm, as the average was found after many experiments, so even if the experiments can be performed a dozen times, it will be found that the average is still 432 mm.

> In order for the average to remain 432 mm, the reading for the sixth time should be 432 mm exactly.

Responses associated with the set paradigm suggested that since the readings are not identical, the true value cannot be selected but lies within the interval given by the readings:

Table 4. Students' reasoning when processing repeated readings ($n = 76$).

	Pre-instruction reasoning (%)			Post-instruction reasoning (%)		
Probe description	Point	Set	Neither	Point	Set	Neither
Representing a set of repeated readings	29 (38)	41 (54)	6 (8)	2 (3)	74 (97)	0 (0)
The meaning of the average value of repeated readings	7 (9)	65 (79)	4 (5)	2 (3)	74 (97)	0 (0)
Predicting a subsequent reading based on the average	3 (4)	68 (89)	5 (7)	1 (1)	75 (99)	0 (0)

> Because all the answers they got were ranging between 426 and 436, so it (the value of d) should be somewhere between that.

> Seeing that the average value is 432 mm, when the ball is rolled again, the reading will either be a little more or a little less than 432 mm.

Post-intervention responses typically alluded to the idea that given a set of repeated readings a standard uncertainty must be calculated to determine the interval within which the measurand may be found:

> The value will be $432.0 \pm u(d)$ if standard uncertainty is calculated which means that it lies somewhere between 431.5 and 432.5 mm.

Table 4 summarizes the responses to these three probes, reflecting students' ideas of measurement when processing multiple observations. When relating a collection of data values to the measurand 38% of the respondents used point reasoning before the course, and only 3% after the course, while most students viewed the average value as the best estimate of the measurand and suggested that a standard uncertainty is a necessary component of the measurement. The other probes did not provide a sensitive differentiation between point and set reasoning before and after the course where point reasoning was identified in less than 10% of the respondents.

5.4. Students' ideas about comparing data sets

Two probes dealt with the comparison of two sets of data resulting from similar experiments. The first probe presents two sets of observations having the same mean but different spreads (ranges), and the data in the second probe have the same spread but slightly differing means. Respondents are asked to comment on the relative agreement and quality of the results.

Typical point reasoning identified in the responses was based on the idea that the results of two sets of readings were equally good since the averages were identical:

> Both sets of results are equally good. As long as the averages are the same, the individual distances do not matter that much.

Similarly, point reasoning was identified when respondents argued that data sets provided incompatible results because their averages differed:

> Their averages are not the same therefore, no matter what the results are, they cannot agree.

Table 5. Students' reasoning when comparing sets of readings ($n = 76$).

	Pre-instruction reasoning (%)			Post-instruction reasoning (%)		
Probe description	Point	Set	Neither	Point	Set	Neither
Comparing quality of sets of readings with same mean and different spreads	54 (71)	18 (24)	4 (5)	26 (34)	49 (64)	1 (1)
Comparing similarity of sets of readings with different mean and the same spreads	65 (85)	6 (8)	5 (7)	18 (23)	57 (75)	1 (1)

Set reasoning in these probes is illustrated by the following examples where the respondents refer to the different spreads (or uncertainties) associated with the different sets of data:

> Even though the average for the two groups are the same, Group B's data values are more dispersed than Group A. This shows that Group A worked more carefully than B.

> Because the intervals of both the students overlap each other if the standard uncertainty for d is calculated therefore making their data values agree.

Table 5 summarizes the responses in terms of the point and set paradigms. Before the course, 24% and 8% of the students used set reasoning when comparing sets of data for the two probes, respectively. The vast majority of students based their comparisons solely on the given averages of the data sets. The post-instruction responses showed that 64% and 75%, respectively, realized the importance of considering the spread in the data, or the standard uncertainty, when comparing sets of results.

5.5. Students' ideas about the nature of uncertainty in measurement

One probe dealt with students' beliefs about whether or not uncertainties associated with measurement can in principle be reduced to zero. The most typical response associated with the point paradigm was that the true value can be found through practice:

> Practice makes perfect, so if the same experiment is done very carefully a perfect reading will be obtained.

On the other hand, set responses typically cited various factors that contribute uncertainties to the measurement result:

> We will never know the true value of d. It doesn't matter how much a human practices the experiment, there are other things in nature that affect it, such as air resistance and a difference in speed due to gravity.

> Because of uncertainties which affect our experiment, like the zero reading, internal calibration, pressure influence and so on.

Table 6 shows that set reasoning was used by 62% of the students before the course, while at the end of the course almost all the students (92%) displayed reasoning associated with the set paradigm to explain their beliefs about the nature of uncertainty.

Table 6. Students' reasoning about the nature of uncertainty ($n = 76$).

	Pre-instruction reasoning (%)			Post-instruction reasoning (%)		
Probe description	Point	Set	Neither	Point	Set	Neither
Understanding uncertainty in measurement	26 (34)	47 (62)	3 (4)	5 (7)	70 (92)	1 (1)

Table 7. Students' reasoning across all probes pre- and post-instruction ($n = 76$).

		Post-instruction			
		Consistent point reasoning (%)	Mixed reasoning (%)	Consistent set reasoning (%)	Total (%)
Pre-instruction	Consistent point reasoning	0 (0)	2 (3)	2 (3)	4 (5)
	Mixed reasoning	0 (0)	24 (32)	44 (57)	68 (89)
	Consistent set reasoning	0 (0)	1 (1)	3 (4)	4 (5)
Total		0 (0)	27 (36)	49 (64)	76 (100)

5.6. Consistency of students' ideas about measurement across the range of measurement activities

The cognitive change effected by the new course was gauged by the degree to which students were able to display reasoning associated with the set paradigm across the range of measurement activities presented in the full set of probes. A student was classified as a consistent set reasoner if his or her responses to both probes comparing data sets, together with at least five of the remaining seven probes, had been classified as compatible with the set paradigm. A student was classified as a consistent point reasoner if the responses to at least seven of the nine probes were classified according to the point paradigm. All other students were grouped in the 'mixed' reasoning category. The pre- and post-test classifications are shown in Table 7, where it can be seen that, upon entry to the course, only 5% of the students displayed reasoning which could be consistently associated with set paradigm, and similarly with the point paradigm. The vast majority (89%) used mixed reasoning when answering the suite of probes. After the new course, 64% of the students gave responses which could be consistently associated with the set paradigm. The results which we have presented are exact for the sample in question. In order to make a more general comparison, we note that the 95% confidence interval associated with this proportion of students ranges from 53% to 74%.

The effectiveness of the new course was gauged by determining the degree to which students were able to provide responses which could be associated with the set paradigm across the full set of probes on completion of the laboratory courses for both the experimental group (using the GUM-advocated approaches for data analysis) and the control group (using traditional approaches for data analysis). Table 8 summarizes the frequencies of student reasoning. It can be seen that after the course featuring traditional approaches to data analysis (the control group), about one in three of the students (38%) used set reasoning consistently, whereas almost two out of three students (64%) reached this stage after the new course based on the GUM framework (the experimental group). It is also notable that at the end of the traditional course, one in eight students (13%) were still consistent point reasoners, while no student displayed this type of reasoning after the new course. A chi-squared calculation for

Table 8. Students' reasoning about scientific measurement for the control group ($n = 70$) and experimental group ($n = 76$).

Student reasoning	After a course featuring traditional data analysis (%)	After the new course featuring GUM-advocated data analysis (%)
Consistent point reasoning	9 (13)	0 (0)
Mixed reasoning	34 (49)	27 (36)
Consistent set reasoning	27 (38)	49 (64)

the significance of the different outcomes of the two courses results in a value of 26.5 (df 2): $p < 0.0001$, confirming the conclusion from inspection of the data in table 8, i.e. the understanding of scientific measurement of students who have completed the course based on the GUM framework (the experimental group) is significantly better than that of students who have been exposed to traditional frequentist approaches for data analysis (the control group).

6. Discussion and conclusion

We have described a new introductory physics laboratory course on measurement [22] that is based on the GUM-advocated probabilistic approach for metrology [18, 20]. This approach offers a consistent method for making inferences about a measurand for both single and multiple observations and introduces unambiguous terminology for communicating measurement results. The course has been evaluated by exploring students' reasoning when processing both single and multiple (repeated) observations, determining the similarity and relative quality of data sets, and explaining the meaning of an uncertainty interval. The evidence strongly suggests that the new course has been significantly more successful in improving students' understanding of measurement and uncertainty than a laboratory course featuring traditional frequentist statistics. Our findings are consistent with research [26] which shows that a teaching sequence based on the principles invoked in the GUM framework improves students' understanding of a measurement based on a single observation.

Recently we have reported on a study of a cohort of students intending to major in physics, thus with a much stronger physics foundation than the experimental and control groups reported on in this paper, and who followed an introductory laboratory course that used the frequentist approach [27] for data analysis. Even for this group of students, with their strong physics and mathematics backgrounds, it was found after instruction that only 19% of the sample could be described as having a good understanding of the nature of measurement uncertainty. This represents a 95% confidence interval of 9–32% which can be compared directly with the present interval of 53–74%. The difference between the results is statistically significant and that the latter gain can also be claimed as having pedagogical significance.

Evaluations of introductory laboratory courses emphasizing the explicit teaching of the concepts underpinning measurement and uncertainty, but based on the frequentist approach, indicated convincing gains in students' understanding of measurement for data collection and data processing [9, 11, 12]. Our evidence suggests that a course using the GUM framework for data analysis also considerably improves understanding of uncertainty. We argue, therefore, that the introductory physics laboratory course should feature activities using the GUM framework as the basis of data analysis and that the experimental aspects of physics be highlighted explicitly rather than relegated to an 'add on' to the theoretical content. By

introducing the concepts of probability and uncertainty as early as possible in the laboratory course, the tentative, yet quantifiable, nature of scientific knowledge is foregrounded, which then allows for meaningful experimentation to take place in any introductory physics laboratory curriculum.

Acknowledgments

We acknowledge the contribution of Bob Campbell to the development of this course and the many people who provided useful comments, including Rebecca Kung and Roger Fearick. This work was partially funded by the National Research Foundation of South Africa. We thank the anonymous referee whose comments considerably improved a previous version of this manuscript.

References

[1] Robinson M C 1979 Undergraduate laboratories in physics: two philosophies *Am. J. Phys.* **47** 859–62
[2] Phillips T 1981 Early history of physics laboratories for students at the college level *Am. J. Phys.* **46** 522–7
[3] Carlson E 1986 Constructing laboratory courses *Am. J. Phys.* **54** 972–6
[4] Schumacher D 2007 Student undergraduate laboratory and project work *Eur. J. Phys.* **28** (editorial)
[5] American Association of Physics Teachers Committee on Laboratories (Gerald Taylor, Jr, Chair) 1998 Goals of the introductory physics laboratory *Am. J. Phys.* **66** 483–5
[6] Tiberghien A, Veillard L, le Marechal J-F, Buty C and Millar R 2001 An analysis of labwork tasks used in science teaching at upper secondary school and university levels in several European countries *Sci. Educ.* **85** 483–508
[7] Séré M-G, Journeaux R and Larcher C 1993 Learning the statistical analysis of measurement error *Int. J. Sci. Educ.* **15** 427–38
[8] Allie S, Buffler A, Kaunda L, Campbell B and Lubben F 1998 First year physics students' perceptions of the quality of experimental measurements *Int. J. Sci. Educ.* **20** 447–59
[9] Buffler A, Allie S, Lubben F and Campbell B 2001 The development of first year physics students' ideas about measurement in terms of point and set paradigms *Int. J. Sci. Educ.* **23** 1137–56
[10] Deardorff D L 2001 Introductory physics students' treatment of measurement uncertainty *Unpublished PhD Thesis* North Carolina State University
[11] Kung R L 2005 Teaching the concept of measurement: an example of a concept-based laboratory course *Am. J. Phys.* **73** 771–7
[12] Etkina E, Murthy S and Zou X 2006 Using introductory labs to engage students in experimental design *Am. J. Phys.* **74** 979–86
[13] von Aufschnaiter C and von Aufschnaiter S 2007 University students' activities, thinking and learning during laboratory work *Eur. J. Phys.* **28** S51–60
[14] Allie S, Buffler A, Campbell B, Lubben F, Evangelinos D, Psillos D and Valassiades O 2003 Teaching measurement in the introductory physics laboratory *Phys. Teach.* **41** 394–401
[15] Bevington P R and Robinson D K 2003 *Data Reduction and Error Analysis* 3rd edn (New York: McGraw-Hill)
[16] Fairbrother R and Hackling M 1997 Is this the right answer? *Int. J. Sci. Educ.* **19** 887–94
[17] Bich W, Cox M G and Harris P M 2006 Evolution of the 'Guide to the Expression of Uncertainty in Measurement' *Metrologia* **43** S161–66
Kacker R, Sommer K-D and Kessel R 2007 Evolution of modern approaches to express uncertainty in measurement *Metrologia* **44** 513–29
[18] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP and OIML 1995 *Guide to the Expression of Uncertainty in Measurement (GUM)* (Geneva: International Organization for Standardization)
[19] Kirkup L 2002 A guide to GUM *Eur. J. Phys.* **23** 483–7
[20] Taylor B N and Kuyatt C E 1994 Guidelines for evaluating and expressing the uncertainty of NIST measurement results *NIST Technical Note 1297* (Gaithersburg, MD: National Institute of Standards and Technology) Available in electronic form at http://physics.nist.gov/Pubs/guidelines/contents.html
[21] Cox R T 1946 Probability, frequency and reasonable expectation *Am. J. Phys.* **14** 1–13
D'Agostini G 1999 Teaching statistics in the physics curriculum: unifying and clarifying the role of subjective probability *Am. J. Phys.* **67** 1260–8
[22] Buffler A, Allie S, Lubben F and Campbell B 2007 *Introduction to Measurement in the Physics Laboratory. A Probabilistic Approach* 3.4 edn (Department of Physics, University of Cape Town) Can be downloaded, and used by instructors, from http://www.phy.uct.ac.za/people/buffler/labmanual.html
[23] The full questionnaires, together with the coding schemes for each probe, can be downloaded from http://www.phy.uct.ac.za/people/buffler/edutools.html

[24] Redish E F 2003 *Teaching Physics with the Physics Suite* (Hoboken, NJ: Wiley)
[25] Lubben F, Buffler A, Allie S and Campbell B 2001 Point and set reasoning in practical science measurement by entrant university freshmen *Sci. Educ.* **85** 311–27
[26] Evangelinos D, Psillos D and Valassiades O 2002 An investigation of teaching and learning about measurement data and their treatment in the introductory physics laboratory *Teaching and Learning in the Science Laboratory* ed D Psillos and H Niederrer (Dordrecht: Kluwer) pp 179–90
[27] Volkwyn T S, Allie S, Buffler A and Lubben F 2008 *Phys. Rev. S. T. Phys. Ed. Res.* **4** 1–10