Rebecca Kung and Cedric Linder work in the Physics Education Research Group that is, for Scandinavia, uniquely situated in a physics and engineering teaching and learning environment (Uppsala University's Ångström laboratory). The group's research focus is on teaching and learning at the university level. One part of this focus is centred on the physics student laboratory – analysing what occurs, what could potentially occur, and socio-cultural dynamics that affect the learning experience. Here, one interest is in how students think about and apply the ideas behind measurement and uncertainty.

REBECCA LIPPMANN KUNG
Department of Physics, Uppsala University, Sweden
Rebecca.kung@fysik.uu.se

CEDRIC LINDER
Department of Physics, Uppsala University, Sweden and Physics Department, University of Western Cape, South Africa
Cedric.Linder@fysik.uu.se

University students' ideas about data processing and data comparison in a physics laboratory course

Abstract
This study investigates undergraduate students' ability to use the ideas of measurement and uncertainty to process and compare experimental data. These ideas include not only knowing what it means to use an instrument to take a measurement, but also being able to apply that knowledge, including the ideas that make up uncertainty analysis, to every aspect of an experiment. A physics laboratory course for the Energy Systems Engineering programme at Uppsala University has been designed to focus on teaching students the ideas of measurement and the associated laboratory skills. In the reported study, we use an open-ended survey to investigate students' ideas about data processing and data comparison before and after this laboratory course. The results show that several students, even after the course, are still unable to appropriately use the ideas of uncertainty. This suggests that these ideas must be continuously revisited and explored as a fundamental part of all undergraduate laboratory experiences.

INTRODUCTION

Data processing and data comparison principally involve the ideas of measurement and uncertainty. Measurement and its related uncertainty are at the very heart of empirical science, and as such are widely considered to be one of the most fundamental and important components of a student's science education (for example, Duggan & Gott, 2002; Welzel et al., 1998). Each phase of an experiment – the design, performance, analysis, and conclusion phases – requires that students know what it means to take a measurement and be able to apply this knowledge along with an understanding of the associated uncertainty. We propose that the underlying ideas for understanding uncertainty can be broadly categorized as follows: (Similar lists may be found in Deardorff, 2001, and Fairbrother & Hackling, 1997.)

- All measurements have an associated uncertainty, which can and should be quantified and reported.
- A calculated result has an associated uncertainty based on the uncertainty carried in its dependent values.

- The design of an experiment and skill of conducting the experiment affects the extent of the uncertainty in a measurement.
- It is impossible to scientifically compare results and draw conclusions without taking into account uncertainty.

These ideas are critical for any experiment to have scientific merit. Of course, throughout their lives people are carrying out "experiments" and drawing conclusions from them, such as trying to find the largest bag of oranges to buy at the grocery store, the fastest driving route, or the best recipe. Such experiments and conclusions may or may not have scientific merit, but they are convincing to the person using them. However, in this article, we are discussing measurements, experiments, and conclusions that are appropriate and informed in a scientific sense.

Some readers may think that understanding measurement is trivial, so this article begins by arguing for the importance and difficulty of understanding measurement, giving examples of the varied situations where measurement is needed. Existing research on students' understanding of measurement uncertainty is given, followed by the description of the context and the open-ended survey used. Finally, the survey results from before and after the laboratory course are described, as well as interpretations of these results and indications for instruction.

Measurement and Uncertainty

Suppose a student performs a common student-laboratory physics experiment: measuring the period of a pendulum to see how it depends upon the length of the pendulum. Without a basic understanding of measurement, the student would be unable to decide:

- How to set up the pendulum – should one tie the string tightly or loosely to the support?
- How to measure the period – should one start and stop timing when the pendulum swings through its lowest point or through its highest point?
- How many measurements to take – how many times a measurement needs to be repeated, and, if necessary, which data points to keep and which to discard.

A student's success with even this very basic experiment thus would depend on their understanding of measurement. (Of course, physics content knowledge is also important to be successful.)

An understanding of measurement is critical for making informed decisions in many different situations – including everyday life situations. For example, suppose a doctor measures a person's blood pressure to be 133/87 mmHg, and the official healthy cut-off is 130/85 mmHg. Whether or not the person should start taking medicine or change their lifestyle depends on many details of the measurement, including the ambient temperature, how quickly the doctor reduced the pressure while taking the measurement, and even whether the person's feet were hanging or resting on the floor (Bickley & Hoekelman, 1999). Or, suppose someone is deciding whether or not to replace their fluorescent light bulbs with full-spectrum lighting – do they simply follow advertisements for full-spectrum lighting quoting studies claiming that it has many health and psychological benefits, or should they think about problems with the measurement of these claimed benefits (such as lack of control, lack of internal validity, and low statistical effects -- see McColl & Veitch, 2001)? In both of these examples a person would need at least a fundamental understanding of measurement to make good sense of the information at hand. Other examples include things like patients being able to appreciate the benefit of a screening mammography (Schwartz, Woloshin, Black & Welch, 1997), doctors being able to determine a patient's treatment (Sheridan & Pignone, 2002), and a person being able to find and hold employment (Bynner, 2004).

The ability to acquire, analyse, and evaluate data has internationally been named as vital for employability and for effective work participation. For example, see the British National Skills Task Force, the U.S. Secretary's Commission on Achieving Necessary Skills, and the Australian Educational Council (as cited in Kearns, 2001). In Sweden, laboratory teaching that includes experimental planning and data analysis has been strongly recommended by the Swedish Högskoleverket (Lindesjöö, 2005). In the EU, many university teachers report that teaching uncertainty is a fundamental aim of their physics laboratory courses (Welzel et al., 1998). To this end, however, research must be done to evaluate the achievement of this goal. Towards this, we use a survey, built on previous research in this area, to investigate how well this goal is achieved.

Previous Research

Existing research in the area focuses on identifying student difficulties and testing students' ideas about measurement. Students' ideas about the reliability and validity of experimental evidence have been studied in the contexts of physics, chemistry, biology, and general science, in several different countries, and at levels from primary to university (see, for example, Evangelinos, Psillos, & Valassiades, 2002; Kanari & Millar, 2004; Masnick & Morris, 2002).

The largest study to compare reasoning across different contexts was done by surveying more than 600 students in six European countries (Leach et al., 1998). Using a written survey consisting of both open-response and closed multiple-choice questions, the researchers found that many students think as follows:

- It is possible to make a perfect measurement (that is, a measurement without any uncertainty) of a quantity given enough time and money (30%-60% of students).
- One should *always* use the arithmetic mean to obtain a final result from a set of data (80% of students).
- The average is all that matters when comparing two data sets, even if they have different confidence intervals (around 30% of students).

Other studies have shown that students rarely *spontaneously* carry out multiple trials unless they suspect a flaw in their first measurement (Séré, Journeaux & Larcher , 1993). In general, students in the laboratory appear to be focused on searching for a specific "true value" without giving due consideration to uncertainty.

The most recent and complete work in student ideas of measurement was carried out by researchers at the Universities of York, UK and Cape Town, South Africa (for a comprehensive summary see Campbell, Lubben, Buffler, & Allie, 2005). Based on previous studies, the researchers developed a framework for interpreting students' reasoning as belonging to what has been characterized as the "set" or "point" paradigms (Buffler, Allie, Lubben & Campbell, 2001). The latest formulation is shown in Table 1 (Buffler, Allie, Lubben & Campbell, 2003, p.2). The researchers further distinguished students' responses by the actions they took and the reasoning behind these actions; for example, students often are able to, and do, calculate an average (a set-like action) but are unable to interpret their calculated average appropriately, using set reasoning.

The researchers used this framework to analyze student responses to a free-response survey they developed, called the *Physics Measurement Questionnaire*. This survey asks students how to best deal with data collection, data processing, and data comparison in a scientific setting. Every question uses the same experimental context: a ball rolling off an elevated ramp onto the floor. (See figure 1.)

The responses from 70 first-year students at the University of Cape Town were analyzed before and after a physics laboratory course. Before instruction, the percentage of students reasoning from the point-paradigm on the five data collection and data processing questions ranged from 54% to 77%. This decreased to 13% to 21% of the students after instruction. However, when asked to compare two data sets of five trials each with the average given, no students were coded as using set-paradigm reasoning, and 98% of the students were coded as giving mixed reasoning. Most of the students answered by comparing the two averages, and they were coded as mixed because they

Table 1: *The set and point paradigm*

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" 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.

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.
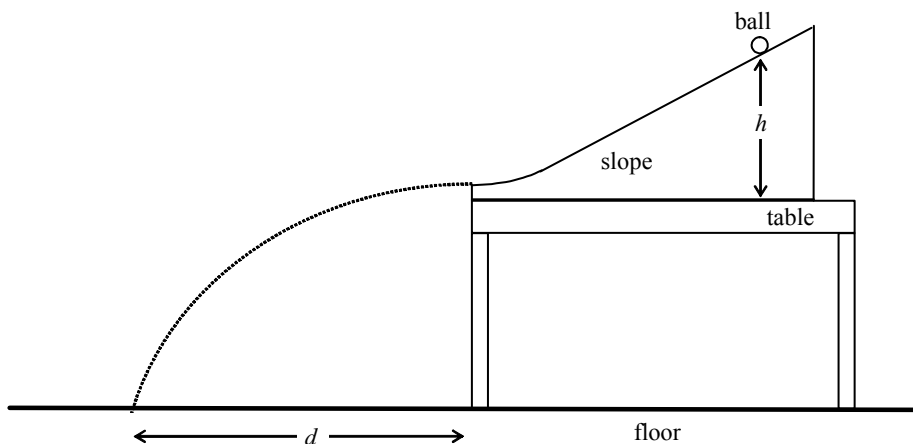

Figure 1. *Task Context for the Physics Measurement Questionnaire (Buffler et al., 2001, p. 1140)*

used the idea of average but did not give any evidence that they considered all the other measurements. These students were able to use the mathematical tools of the set-paradigm, but were unable to support the tools with reasoning based on an understanding of the set-paradigm. Buffler et al.'s instructional goal is to enable students to move toward reasoning using the set-paradigm, not just using the mathematical tools.

We agree that students must develop an understanding of the underlying ideas – not just build competence with the calculations. Different fields have different methods for calculating, reporting, and comparing uncertainty. Even among physicists there is much variation in the expression of uncertainty (Deardorff, 2001). Thus teaching students one specific method for calculating and expressing uncertainty may not be useful. However, the ideas of measurement uncertainty are both necessary and applicable in any scientific domain.

The study

Even though a base of research on students' ideas of measurement uncertainty exists, few studies look at how a course changes students' ideas (see Abbott, 2003; Buffler et al, 2001; Rollnick, Lubben, Lotz & Dlamini, 2002; Séré et al, 1993). These limited studies have all taken place in contexts widely different from Nordic universities, and have focussed on students under-prepared for the science laboratory, first year students, or non-science majors. Hence, in this study we investigate students in their second year of an engineering program at Uppsala University, analyzing their ideas about measurement through an open-ended survey. The research questions may be delineated as:

1. What do the students reveal about their understanding of measurement when asked to report, compare, and combine measurement data?
2. After students have taken the Energy Systems university physics-laboratory course, is any meaningful re-constitution of these ideas evident?

Research Context

The students who participated in this study were enrolled in the *Energy Systems Engineering Programme (ES)* at Uppsala University. The associated mechanics physics course, given in the second year of the programme, is the first physics course these students take at university. The course consists of 5 weeks of full-time study spread over the 20 week semester. The laboratory component of the course accounts for one week of full-time study (1/5 of the entire course). The students are reasonably comfortable and competent with mathematics and science, having already taken linear algebra, multi-variable calculus, biology, and chemistry in their first year of coursework.

The ES mechanics course laboratory was redesigned in 2004 based on the *Scientific Community Laboratory*, which aims at creating a natural scientific community for developing students' ideas about measurement and uncertainty (Kung, 2005). The course consists of 6 two-hour laboratory exercises and 1 four-hour project laboratory. During the two-hour laboratories the students are given a one-page handout describing the laboratory question. The students work in groups of four to design a method, perform the experiment, analyse the data, form a conclusion, and present this to the class. The laboratory questions are open-ended and designed to promote students' engagement with specific measurement ideas. Frequently, a theoretical answer for an ideal situation is known, but the answer for the actual laboratory situation is not known. Each laboratory exercise can be answered using several different experimental designs, encouraging students to critique different experimental methods. The last 20 minutes of the laboratory is spent in a class discussion, where each group presents their method and conclusion, with the rest of the students asking questions. The aim of the discussion is to compare the methods and analyses to determine convincing scientific ways to take, analyse, and present data.

For example, suppose a group of students measured the force of friction on an object by placing the object on a board and tilting the board at increasing angles until the object began to slide. During the class discussion, the students present the average largest angle for two objects with the same material and mass but different contact area: 30.1 degrees and 34.8 degrees. The students claim that this data shows that a different contact area causes a different friction force. Several students in the audience would probably disagree with this claim, and say that 4.7 degrees is not a significant difference. Through further discussion the students would be encouraged to present their full data, give some measure of the uncertainty, and use the uncertainty to argue whether 4.7 degrees is significant or not.

Design of the survey

The survey used in this study was built upon the *Physics Measurement Questionnaire* described earlier. However, the physics measurement questionnaire was designed for a different student population – students severely under-prepared in science, particularly in the student laboratory. Previous experience with the questionnaire, including use at the University of Maryland, USA (Lippmann, 2003) and a trial in Uppsala during 2004, had revealed which questions were appropriate for the Uppsala student population, and also prompted the adding of certain, more difficult, questions. (See the Appendix for survey details.)

The final survey, given in Autumn 2005, consists of 7 free-response questions in Swedish, asking students:

- how to report data;
- how to compare data quality;
- the number of trials needed to compare data quality;
- whether two sets of data agree or disagree; and,
- how to combine two sets of data.

METHOD

Students took 20-30 minutes to complete the survey during both their first laboratory and their final project laboratory meeting. There were four weeks of classes in between the pre-survey and the post-survey. Students were told that the survey would be used to evaluate the course and its outcomes, and for this purpose their names would be used only to allow before/after comparisons to be made. Then, only students who completed both before and after sets of surveys were included: a total of 41 students. The students' responses were read, interpreted, and grouped with other similar responses. As an example, consider the answers to the first question, which shows students the data for five trials and asks the students to describe what should be reported for the final result. Most students answered that they should report the average. Some students answered that they should eliminate the highest and lowest value and average the rest. Some students added that the spread in the data should be reported as a number, or that a diagram should be drawn to show the spread. The number of students in each category were counted and compared by one researcher. Two researchers individually coded 10% of the surveys and compared their results. Coder agreement was 93% before discussion and 100% after discussion. Data is given in number of students (out of 41).

SURVEY RESULTS

The results from the survey showed some large changes from pre to post, and also some interesting aspects that did not change. These results are grouped by question topic, and reported here. Most quoted responses have been translated from Swedish into English by the first author. A few quotes

were originally written in English – these are indicated. (The course was given using both Swedish and English). Ideal answers have also been included in an Appendix.

For each question, students' answers could change from pre-test to post-test – moving to a more appropriate response or moving to a less appropriate response. Students could also keep their less appropriate or more appropriate response the same. The numbers of students in each of these four categories are discussed.

Reporting uncertainty

Students were given several opportunities to decide how to report experimental results. Ideally, the students would describe how reporting just an average, for example, leaves out information about the spread in the measurement. Students may use standard deviation, maximum-minimum range, or any number of other calculations to report the spread in the data. For the first question, 13 students switched from not using spread in the pre-test to using spread in the post-test, and three students changed in the other direction. Thus a total of 18 students included some form of spread in the pre-test, and 28 in the post-test. For example, one student answered "The range is good to show how much the separate values differ from each other. The average is good to show the size of the values." This result does show an improvement in whether students appreciate reporting the uncertainty in a measurement, but a rather disappointing improvement, since 10 students are still lacking this most basic reasoning (in both pre and post tests). This becomes even clearer in a question given to the students as the laboratory part of their final examination. Here a similar question was asked: "What should the student report to the neighbouring group so they can compare their answers?" One student's response gave only the average and said (in English) "The standard uncertainty gives an idea of what the value could range to and from but is not essential in a comparison." This answer is the exact opposite of what was wanted – that without the uncertainty there is no way to compare results. These students are showing a belief, apparently resistant to change, that one can compare experimental results just by comparing the two averages and seeing whether they are the same or not.

Question 4 of our survey asked students to combine two sets of data, given the data points and the average. For this question, nine students switched from not reporting uncertainty in the pre-test to reporting uncertainty in the post-test, and two students changed in the other direction. Thus a total of nine students reported uncertainty in the pre-test, and 16 in the post-test. In this question, students were focusing on how to combine results, and not on how to appropriately report results. However, they should still have been aware that reporting only an average without uncertainty means ignoring essential information. Similarly, in question 6, students were asked to combine two sets of data, given the average and the standard deviation. For this question, the total number of students reporting the uncertainty did not change significantly – 22 in the pre-test, 21 in the post-test. Seven students changed from reporting uncertainty in the pre-test to not reporting uncertainty in the post-test, and 6 students changed in the other direction. A larger number of students report uncertainty for this question because they are provided with an uncertainty in the question text, and repeat that format (average ± uncertainty) in their response. Several students appear to be repeating that format without realising the significance of it, as shown by their willingness to change between the pre and post test. It is disappointing that the course did not encourage more students to better understand the meaning and importance of the format, and thus repeat it in their answer.

Using uncertainty to compare data quality

In question 2, students were shown 5 trials and the average for two groups, and asked which group had the best result. The averages were the same, but the spread in the data was different. For this question, two students switched from claiming the data with the smaller spread was better in the pre-test to saying they were the same in the post-test, and seven students switched in the other

direction. Thus, a total of 34 students mentioned that the smaller spread data was better in the pre-test, and 39 students in the post-test. This appears to be a context where students have less difficulty being able to use uncertainty.

Determining quantity of data

Students were also asked in question 2 to write down how many trials are needed to decide which group's results are better. In the pre-test, students answered ranging from 2 to 20 trials giving an average of 4.4 trials with a standard deviation of 4.3. In the post-test, students answered ranging from 2 to 20 trials giving an average of 6.3 trials with a standard deviation of 4.7. (These numbers disregard one student's answer of 50. Including the 50, the average of the student's answers is 7.7, and the standard deviation is 9.0.) The number of students answering that 3 trials is enough dropped from 16 in the pre-test to 8 in the post-test. Based on experience teaching the course, it appears that many students enter university with a belief that three measurements is a good number. The laboratory course may have alerted the students to situations where three measurements are insufficient.

Comparing results

Questions 3, 5, and 7 asked students to evaluate the agreement or disagreement of two groups' experimental results. Students could either use the uncertainty to make their decision (by looking to see if the two ranges overlap, or comparing the difference in the averages to the uncertainty), or ignore the uncertainty and make a judgement of feeling by comparing numbers (typically the average). For example, one student wrote, "The intervals containing the values overlap, so you can say that they agree." In contrast, a different student wrote that the results "probably do not agree because they got different averages." This student did not use the uncertainty to decide whether the difference between the averages is significant or not. A third student did not consider any aspect of the data to check whether they agree. This student wrote "They have performed the same experiment and they have done the same number of measurements, therefore they have produced the same result."

Question 3 gave students the data points and the average for both data sets. For this question, 5 students used uncertainty in the pre-test and 30 students in the post-test. Questions 5 and 7 gave students the average and standard uncertainty for both data sets. The number of students using uncertainty increased from 22 to 36 for question 5, and 23 to 31 for question 7. (For these three questions, only one or two students who answered using uncertainty in the pre-test failed to use uncertainty in the post-test.) Similarly to when students are asked to report data, seeing uncertainty in the question text makes it more likely that they will use uncertainty in their response. However, a similar number of students used uncertainty in the post-test. A possible interpretation of this is that a certain group of students were able to use uncertainty when prompted, before the course. These students became able to use uncertainty after the course without prompting. There also was a group of students whose learning did not seem to be influenced by the course. These students were unable to use uncertainty with or without prompting, before or after the course.

Using diagrams

When students reported their results during the class discussion, they were encouraged to use diagrams, such as histograms or linear scatter plots. This seemed to impact some of the students, as it showed up on the survey. Even though no diagrams were used in the survey, 11 students answered on the post-test that a diagram should be included in the report for question 1. No students had mentioned a diagram in the pre-test. It appears that some students found diagrams useful in their laboratory work, and decided that they should be used when reporting data.

Discussion

The survey results do show an interesting pedagogical difference in how students process and compare data from pre-test to post-test. Much of this is encouraging, especially considering the lack of change demonstrated in a traditional laboratory course (Abbott, 2003). Unfortunately, there were still a number of students who did not improve their understanding of the importance of reporting uncertainty and/or using uncertainty to determine agreement between sets of data.

One could argue that this is a difficulty the students have with the statistical calculations – they need more practice finding the average, standard deviation of the mean, coverage probability, etc. However, the students showed that they had the ability to do these calculations during the course. The difficulty comes when they need to use uncertainty to interpret the results of a measurement. The difficulty is not with the calculations, but with the knowledge that, and how, uncertainty concepts should be applied to all measurements.

It appears challenging to shift students from the earlier discussed point-paradigm to the set-paradigm approach. Many continue to see measurement as being about the search for one correct answer (for these students, the average) rather than finding the best approximation, with a certain probability of being larger or smaller. The students are able, however, to judge the quality of an experiment from the range of the data. It seems a small step to move from "smaller range means better result" to "better result means better known average." The next step would be to realise that how well one knows the average influences whether two averages agree or disagree. This appears to be the most difficult step. The next design of this laboratory course, to be taught in 2006, will attempt to promote this sequence of reasoning. The plan is that students will first be asked, during the class discussion, to consider which method produced better results, then how well the average is known for two different methods, and then whether the averages agree or disagree.

The students investigated in this study are second year students in a typical Swedish *civilingenjör* programme. Such students are experienced in the student laboratory, having focused on science in secondary school and in their first year and a half at university. They are able to judge the quality of an experiment based on the range of data – as opposed to less-experienced student populations mentioned earlier. Yet, these students are still showing difficulty understanding measurement. This sort of student population – students with significant science laboratory experience – is common in Nordic countries, and we believe these results will also apply to students in similar programs in other Nordic universities. In our experience such students most often are taught uncertainty analysis in a brief setting with a focus on the mathematical calculations, and it appears that such education does not promote the development of even a fundamentally appropriate understanding of uncertainty. However, in our conversations with colleagues it never ceases to surprise them that such students continue to have difficulties with these basic ideas.

Conclusion

Understanding measurement and its related uncertainty is critical for those engaging in experimental research, interpreting experimental results, and making informed decisions in everyday life. This paper focuses on how students use the ideas of measurement and uncertainty to process and compare experimental data, showing that these ideas are not necessarily understood as they should be even by university science students in their second year. For example, 11 out of 41 students failed to apply the basic idea that uncertainty must be used to compare the results of two sets of data, even after a specially-designed laboratory course. It appears difficult to adequately promote an appropriate understanding of measurement even through a specially-designed laboratory course. This contradicts a frequently-heard opinion that one laboratory exercise is sufficient to teach uncertainty effectively. We thus recommend that these ideas be continuously revisited and systematically explored as a fundamental part of all undergraduate laboratory experience.

Acknowledgements

References

Abbott, D. S. (2003). *Assessing student understanding of measurement and uncertainty*, unpublished doctoral dissertation, North Carolina State University.

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

Bickley, L. S., & Hoekelman, R. A. (1999). *Bates' guide to physical examination and history taking* (7th ed.). Philadelphia: Lippincott.

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

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

Bynner, J. (2004). Literacy, numeracy and employability: evidence from the British cohort studies. *Literacy and Numeracy Studies*, *13*(1).

Campbell, B., Lubben, F., Buffler, A., & Allie, S. (2005). *Teaching Scientific Measurement at University: Understanding Students' Ideas and Laboratory Curriculum Reform*, Monograph of the African Journal of Research in Mathematics, Science and Mathematics Education. Kloof, South Africa: SAARMSTE.

Duggan, S., & Gott, R. (2002). What sort of science education do we really need? *International Journal of Science Education*, *24*(7), 661-679.

Deardorff, D. L. (2001). *Introductory physics students' treatment of measurement uncertainty,* unpublished doctoral dissertation, North Carolina State University.

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

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

Kearns, P. (2001). *Generic skills for the new economy - review of research*. Adelaide: NCVER.

Kanari, Z., & Millar, R. (2004). Reasoning from Data: How Students Collect and Interpret Data in Science Investigations. *Journal of Research in Science Teaching*, *41*(7), 748-769.

Kung, R. L. (2005). Teaching the concepts of measurement: An example of a concept-based laboratory course. *American Journal of Physics, 73*(8), 771-777.

Leach, J., Millar, R., Ryder, J., Sere, M-G., Hammelev, D., Niedderer, H., & Tselfes, V. (1998). Students' images of science as they relate to labwork learning. *Labwork in Science Education*, Working Paper 6, Targeted Socio-Economic Research Programme Project PL 95-2005.

Lindesjöö, E. (2005). Utvärdering av astronomi, fysik, hydrologi och meteorologi vid svenska universitet och högskolor. Högskoleverket rapportserie 2005:15R: Högskoleverkets vaktmästeri: Stockholm.

Lippmann, R. F. (2003). *Students' Understanding of Measurement and Uncertainty in the Physics Laboratory: Social construction, underlying concepts, and quantitative analysis*. Unpublished Doctoral Dissertation, University of Maryland, College Park.

Masnick, A., & Morris, B. (2002). Reasoning from Data: The Effect of Sample Size and Variability on Children's and Adults' Conclusions. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 643-648). Mahwah, NJ: Lawrence Erlbaum.

McColl, S. L., & Veitch, J. A. (2001). Full-Spectrum Fluorescent Lighting: A Review of Its Effects on Physiology and Health. *Psychological Medicine, 31*(6), 949-964.

Rollnick, M., Lubben, F., Lotz, S., & Dlamini, B. (2002). What do underprepared students learn about measurement from introductory laboratory work? *Research in Science Education. 32*(1), 1-18.

Schwartz, L. M., Woloshin, S., Black, W., & Welch, H. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*, 966-972.

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

Sheridan, S., & Pignone, M. (2002). Numeracy and the medical student's ability to interpret data. *Effective Clinical Practice, Jan-Feb 5*(1), 35-40.

Welzel, M., Haller, K., Bandiera, M., Hammelev, D., Koumaras, P., Niedderer, H., et al. (1998). Teachers' objectives for labwork. Research tool and cross country results. *Labwork in Science Education*, Working Paper 6, Targeted Socio-Economic Research Programme Project PL 95-2005.

APPENDIX

This appendix contains the questions from the survey, with the survey instructions and context removed. For each question, one example of an answer that would be appropriate to expect from a second year university student is given – that is, a point argued using the ideas of measurement and uncertainty. There are multiple ways to appropriately answer each question, some of which may contradict each other, since it depends on the chosen level of confidence. Since the question asks for a decision, the answers also argue a particular decision, demonstrating how to use measurement uncertainty to do this. (These answers do not show a fully scientific answer. One cannot claim that yes, the two group's data agree – one can only claim, for example, that the two groups' data agrees at the 68% confidence level.) The survey has been translated into English to be available to a larger number of readers. A copy of the original survey, in Swedish, is available by emailing the first author.

Some of these questions come directly from the Physics Measurement Questionnaire (see Appendix 1 in Campbell et al, 2005): question 1 (PMQ probe UR), the beginning of question 2 (PMQ probe SMDS), question 3 (DMOS), and question 5 (similar to DMSU). Other questions in the PMQ (such as questions asking whether to perform a second trial or stop with one trial) were shown not to distinguish among students who are experienced in the student laboratory at the secondary and introductory university level. The questions on how much data is enough and on combining data originated from the laboratory class, after students raised the issues. Reducing the number of questions and giving the survey in Swedish increased significantly the number of complete responses collected.

1. One group of students releases the ball down the slope at a height $h = 400$ mm. Their results after five releases are:

Release:	1	2	3	4	5
d (mm):	436	426	438	426	434

They want to present their results to the class. Write down what you think the students should report for their final result for d.

They need to report the average and the complete (max-min) range.
The range shows how much the separate values differ from each other,
and the average shows the size of the values. A histogram might be
good to show more specifically what the data looks like.

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

Based on the five releases, how do their results compare?

_____**Group A's results are just as good as Group B's.**

_____**Group A's results are better.**

_____**Group B's results are better.**

How much data do they need to decide whose results are better? Write down the number of releases needed.

_____ **releases**

Explain how they will know when they have enough data.

> *After 5 trials, the (max-min) range for A is 20 mm, and the range for B is 36 mm. But this range could be low probability (just luck). Perhaps we should keep taking data until the range has stabilized – all the data points fit into the range already established. Or perhaps we should keep taking data until we get a histogram that is starting to form a peak, instead of just being scattered points. It all depends on how sure we want to be that one group is better.*

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

Do the results of these two groups agree?

_____**Yes**

_____**No**

> *Group A ranges from 424-444, and group B ranges from 438-462. Each group has two data points that fit within the other group's data. With the results given, that means 4/10 points overlap. I don't think that is enough to show agreement. It is also hard to tell with only five data points.*

4. The two groups want to combine their results.
Write down what, if anything, you think they should report as their combined result.

> *I have decided that their data does not agree, so they should not combine their results. (Perhaps one group consistently gave their marble a push when they released it.) If I were forced to combine them, I would put all the data points together, and calculate a new average and (max-min) range.*

5. 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 = 434 ± 5 mm

Group B: d = 442 ± 6 mm

Do the results of these two groups agree?

_____**Yes**

_____**No**

> *Using the standard deviation, that means group A's data ranges from 428-439 mm and group B's data ranges from 436-448 mm. This is not a (max-min) range – it is based on the standard deviation, the average distance the data is from the average. These ranges overlap, which means part of the actual data points overlap. I would say these data sets agree.*

6. The two groups want to combine their results.
Write down what, if anything, you think they should report as their combined result.

> *I would put all the original data points together, and calculate a new average and standard deviation. If I did not have the original data points, I would average the two averages, to get 438. Together, the data sets have a lower point of 428 and a higher point of 448, so I would report 438 ± 10 mm.*

7. 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.
Grupp A: d = 434 ± 5 mm
Grupp B: d = 448 ± 5 mm

Do the results of these two groups agree?

_____**Yes**

_____**No**

> *Here we have a standard deviation range of 429-439 and 443-453 mm. There is a 4 mm gap between these ranges, large compared to the standard deviation of 5 mm. I would say that these do not agree.*