Education and Science

Vol 44 (2019) No 199 103-126

Pre-service Physics Teachers' Difficulties in Evaluating Experimental Evidence

Olga Gkioka [1]

Abstract

The reported research is about the difficulties that pre-service physics teachers experience when they judge the quality of experimental results to draw conclusions. Thirty-six pre-service physics teachers participated in the study, enrolled in six semesters (six participants in each semester). They designed, conducted experiments, analyzed the results, evaluated the quality of experimental evidence and finally, evaluated the whole experimental procedure. In addition, the participants were provided with experimental results not collected by themselves (secondary sources data) and were asked to judge the quality of collected evidence and conclusions. Data sources include laboratory reports for each experiment and exam papers, supplemented with individual interviews. A qualitative and quantitative data analysis identified trends within the participants. Findings show that pre-service physics teachers have difficulties with the concepts of experimental validity, measurement reliability, accuracy and precision. Furthermore, they have difficulties in putting such concepts together to judge how well they can rely upon evidence to draw conclusions. The participants also demonstrate difficulties with sources of experimental errors and do not make the distinction between random and systematic errors. In particular, they are confused when they reason about what it is that "gets better" when one takes repeated measurements. It is argued that there is a need for specialized programmes of teacher education to address the development of laboratory skills (i.e. evaluation of experimental evidence) so that preservice teachers become confident in teaching in the physics laboratory in secondary schools. Implications for teaching practice, curriculum development and further research have been discussed.

Keywords

Evaluation of experimental evidence

Validity

Accuracy

Reliability

Physics laboratory

Physics teacher education

Article Info

Received: 07.09.2018

Accepted: 02.28.2019

Online Published: 07.12.2019

DOI: 10.15390/EB.2019.8030

[1] Boğaziçi University, Faculty of Education, Department of Mathematics and Science Education, Turkey, olga.gkioka@boun.edu.tr

Introduction

According to many school science curricula, secondary students should know how to evaluate their own experiments and secondary sources experimental evidence. That is, secondary students should demonstrate a good understanding of reliability and validity of scientific evidence. Measurement reliability and experimental validity are the most common ideas in the secondary school physics laboratory. Such understandings related to the evaluation and laboratory skills are usually under-represented due to the prevalence of confirmatory laboratory courses in many secondary and post-secondary science programs. This results into students' inability to critically evaluate scientific claims (Albers, Rollnick, & Lubben, 2008; Leach, 1999).

In this research study, the aim has been to explore pre-service teachers' understandings closely related to the quality of experimental evidence. Findings from this research will guide the development of a course for pre-service physics teacher preparation.

The undergraduate years are a turning point in producing scientifically literate citizens and future scientists. We want secondary school students to gain experience and confidence in the lab so that they are prepared for employment and higher education. In their turn, science teachers play an important role in preparing scientifically literate citizens and competent candidates to enter the science departments. Science teachers are responsible for the preparation of the next generation of scientists. Well-prepared teachers are of critical importance for student learning and achievement (National Research Council, 2001).

The research discussed in this paper, was influenced by Gott and Duggan's (1995, 1996) position that students, when in the laboratory, call upon a distinct set of conceptions regarding the reliability and validity of scientific evidence when designing and conducting experiments and when evaluating the resulting evidence and conclusions. In simpler words, when students are asked to judge the quality of experimental evidence or evaluate their own experiment, they should deal with the two following issues:

- Can the data be believed?
- Can the data answer the questions?

According to Gott and Duggan (1995, 1996), experimenters need to demonstrate related understandings, which are necessary to explore scientific evidence thoroughly. The same researchers argued that decisions in the lab should be underligned "by 'the thinking behind the doing' of science and include concepts such as deciding how many measurements to take, over what range and with what sample, how to interpret the pattern in the resulting data and how to evaluate the whole task" (p. 186). The same arguments was furthered by Roberts and Johnson (2015) who argued in favour of the 'thinking behind the doing' in scientific practice.

Conceptual and Theoretical Background of the Study

In the physics laboratory, when one evaluates experimental evidence, s/he should consider important concepts like the validity of an experiment and the reliability of measurement. Experimental validity incorporates suitable equipment, identification of variables and appopriate measuring procedures. Reliability of measurements refers to the consistency of measurement procedure (Taylor, 1997). A reliable experiment has results which can be obtained consistently. To ensure that the results are reliable, the experiment must be repeated and consistent results to be obtained (within an acceptable range error). As Lewin and Goldstein (2012) wrote: "The measurement that you make without the knowledge of uncertainty is meaningless". When the random uncertainties in an experiment are small, the experiment is called precise. Precision tells how sure you are of your measurements, regardless of whether your measurement is accurate or not (Bevington & Robinson, 2003; Taylor, 1997). When the systematic uncertainties in an experiment are small, the experiment is accurate. Accuracy is a measure of how close one value is to the accepted value.

Evaluation includes a critical appraisal of the reliability and validity of the experimental procedures being followed. Evaluation of experimental evidence and procedure should be based on the knowledge of what 'counts' as evidence. There are 'rules' in science that help to understand what 'counts' as evidence and lay the foundation for 'good' quality of evaluations. In the evaluation section of a physics lab report, the key main aspects that should be considered and included are, as follows:

- To look critically at the results to judge how well one can rely upon evidence from the experiment to draw conclusions.
- To comment on whether the results are accurate, supporting this with specific data, e.g., % error calculations.
- To comment on whether results are valid and reliable - could other factors have affected the results, would this have been significant, is there any evidence for this?
- To identify and account for any anomalous results (due to inaccuracies in the procedure and/or measurements).
- To talk about limitations in the experiment and suggest possible improvements (specific detail required).
- To talk about specific proposals for further work that would improve or add to the evidence.

Central to the development of this study is the recommendations from three major research reports for the teaching in the laboratory in the USA. Firstly, in the America's Lab Report entitled: "Investigations in High School Science", it was underlined that the "Pre-service education for science teachers often does not directly address laboratory experiences and does not provide teachers with the knowledge and skills needed to lead laboratory experiences" (Singer, Hilton, & Schweingruber, 2005, p. 1). Also, "undergraduate education of future high school science teachers does not currently prepare them with the pedagogical and science content knowledge required to carry out such teaching strategies" (Singer et al., 2005, pp 5-6). In the same report, it was concluded: "Improving high school science teachers' capacity to lead laboratory experiences effectively is critical to advancing the educational goals of these experiences. This could require major changes in undergraduate science education, including providing a range of effective laboratory experiences for future teachers and developing more comprehensive systems of support for teachers" (Singer et al., 2005, p. 216).

Secondly, in the Report "Taking Science to School", it was emphasized that "instruction needs to build incrementally toward more sophisticated understanding and scientific practices. (…) Practices can be supported with explicit structure, or by providing criteria that help guide the work". (Duschl, Schweingruber, & Shouse, 2007, p. 251).

Thirdly, according to the American Association of Physics Teachers (AAPT) (2014, 2017) Reports for the undergraduate physics laboratory, the aim is for secondary school and undergraduate students to be able to design experiments, develop technical and practical skills, analyze and visualize data and communicate physics. Also, undergraduate physics students need to learn how different measurement procedures result in different uncertainties, design improvements to measurements, learn to break down components of experimental design, design experiments to test assumptions and understand limitations of measurement instruments. In addition, students need to know how to communicate their methods, results and findings, with an emphasis on the "why". The American Association of Physics Teachers (AAPT) recommended that explicit instruction should help students reflect on and interpret scientific evidence.

However, in physics departments, as McDermott (1990) explained, it is often assumed that, the participation of physics students in regular undergraduate science laboratory courses provides them with the required knowledge and skills to teach in the physics laboratory. This assumption stands in parallel with the belief that the more rigid and difficult modules one attends in undergraduate programs, the better and well-prepared physics teacher will be (McDermott, 2014).

Yet, little is known about pre-service physics teachers' understanding of experimental evidence. Research needs to be undertaken to look at how and with what criteria preservice physics teachers judge the quality of experimental results and evidence because such an issue did not receive much attention in research.

Research Into Students' Difficulties When Evaluating Experimental Evidence

Many undergraduate physics and secondary school students experience difficulties in the laboratory. The literature provides a range of difficulties encountered and common mistakes made in the physics laboratory. In particular, Boudreaux, Shaffer, Heron, and McDermott (2008) investigated the ability of undergraduate students to reason on the basis of the control of variables. The participants were undergraduate physics students and asked to decide whether a given variable influences the behavior of a system. They found out that although most of them recognized the need to control variables, many had significant difficulty with the underlying reasoning. More studies provide a range of misconceptions related to experimental validity showing that many students do not possess appropriate conceptions of controlled experimentation. Findings of research studies suggest that secondary school and undergraduate students often hold limited understandings of the experimental validity of scientific evidence (Leach, 1999).

Lubben and Millar (1996) investigated children's understandings of reliability of experimental data, calculation of error and measurement error by using a written survey instrument. More than 1000 students aged 11, 14 and 16 years old children participated and were asked to make judgements about the importance of repeating measurements and about the interpretations of data variability. A range of ideas about the function of repeat measurements, how to handle repeat measurements and anomalous readings and the significance of the spread of a set of repeated measurements emerged.

Studies on performance on lab work questioned the assumption that undergraduate students understand the nature of measurement and experimentation (Lubben, Campbell, Buffler, & Allie, 2001; Séré, Journeaux, & Larcher, 1993). After completing a traditional laboratory course, many undergraduate students have ideas about measurement that are inconsistent with the generally accepted scientific model. For example, as Séré et al. (1993) found, a large proportion of undergraduate students view the ideal outcome of a single measurement as an "exact" or "point-like" value. Research into the experimental procedures used by French physics undergraduate students (freshmen) with regard to data handling (Séré et al., 1993) showed that many students hold beliefs about the existence of a single "true" value for a measurement and therefore, attribute any variation in repeated measurements to mistakes. Séré et al. (1993) also concluded that even the correct use of statistical procedures seldom indicates an appreciation of the purposes behind such procedures, or an understanding of how to assess the reliability of data. Only if a measurement is considered really "bad", would it then be reported in terms of an interval (Lubben et al., 2001).

Two types of reasoning used by students in the laboratory, via point and set reasoning, were proposed and used to classify students' responses. Point reasoning is characterized by the underlying notion that each measurement could be in principle be the true value. As a consequence, a measurement is perceived in leading to a single "point-like" value rather than establishing an interval. This way of thinking manifests itself in the belief that only one single measurement is required to establish the true value, as indicated in the work of Séré et al. (1993). Set reasoning is characterized by the notion that each measurement is only an approximation to the true value. As a consequence, a number of measurements are required to form a distribution that clusters around some particular value.

Estimating measurement uncertainties is important for experimental scientific work. However, this is very often neglected in school curricula and teaching practice (Priemer & Hellwig, 2018). Research findings suggest that undergraduate physics and secondary students do not always properly understand, for example, the need for repeated measurements (Grant, 2011, Lubben & Millar, 1996**)**. If students take a series of measurements for any reason, or they are shown a series of measurements, they select a recurring value or the one-to-one comparison of data values. They hold various misconceptions of repeated trials.

Secondary students and undergraduate physics students hold various understandings of the collection and evaluation of experimental data. For example, Allie, Buffler, Kaunda, Campbell, and Lubben (1998) found that undergraduate physics students demonstrated various strategies to handle an anomalous data point. They found out that approximately half of the undergraduate physics students considered variance in their evaluation of the reliability of data. Other researchers observed students' difficulties with the notion of the mean as a representation of multiple measurements (Leach, 1999). Secondary and undergraduate students often hold naïve notions of the rationale of repeated trials, appropriate treatment of anomalies, and the role of variance in establishing reliability. The findings range from primary level (Varelas, 1997), middle and high school (Foulds, Gott, & Feasey, 1992) and in undergraduate science courses (Allie et al., 1998). What is more worrying is that many undergraduate physics students seem to show similar misunderstandings related to the collection and evaluation of experimental data (Gott & Duggan, 1995). "Ideas about data evaluation are particularly poor" (Gott & Duggan, 1995, p 84).

A report on a small-scale study that explored university staff views on laboratory skills in new undergraduates within the Russell Group Universities (i.e. the top 24 research Universities in the UK) also concluded that students were commencing university, lacking not only appropriate skills but also the confidence to carry out practical work within a laboratory (Grant, 2011). Both Grant (2011) and Gatsby (2012) found out that practical skills had declined over the last years and that a factor in the lack of practical skills was the 'limited exposure to practical work at school' (Gatsby, 2012, p. 2).

In France, Caussarieu and Tiberghien (2017) explored how a first-year university physics course deals with measurement uncertainties in the light of an epistemological analysis of measurement. They found that the instructors' expectations are that students systematically estimate uncertainties so that they become aware that measurements and calculations are never exact. However, since uncertainties are not specified for the values given in the laboratory guides, uncertainties are often missing from the results of students' calculations. Another study, carried out in Spain by Crujeiras-Pérez and Jiménez-Aleixandre (2017), explored students' interpretation of data, in particular anomalous results generated by them when carrying out experiments and their ability to monitor them. Such study revealed understandings related to the identification of anomalous points and explanation of sources of errors. The literature reports that many students have difficulties in understanding errors (Lippmann Kung, 2005; Lubben & Millar, 1996; Séré et al., 1993).

At the University of Duisburg-Essen, a very recent study by Eshach and Kukliansky (2018) has showed that students with less laboratory experience use intuitive rules more frequently than students with more laboratory experience. They concluded that understanding the influence of intuitive rules on students' performance, when dealing with experimental data, may be a great help to educators in designing better learning environments to address related science practices. Many scholars like Kalthoff, Theyssen, and Schreiber (2018) have argued in favour of various more or less explicit instructional approaches for the promotion of experimental skills in the training of school teacher students.

Method

The research question guiding the reported study was: "*What are the difficulties that pre-service physics teachers experience when they evaluate experimental results and the whole experimental procedure?"*

The Context of the Study

The study took place within the context of the course "Secondary Physics Lab Applications" of the pre-service teacher education program in a Department of Physics Teaching. It is taught for five hours per week for thirteen (13) weeks a semester. Thirty-six (36) pre-service physics teachers participated in the study registered in six semesters (six participants in each semester). Thus, the participants were the undergraduate students who registered for the course. No selection or changes were made to the students who registered for the course. Thus, such study included convenience sampling. All of them had completed four undergraduate compulsory laboratory courses in the Physics Department.

During the course, the participants revised the concepts of precision, accuracy, validity and reliability of evidence. They also carried out physics experiments without being provided with detailed guidelines. Indeed, they performed experiments starting from revising the relevant theory and deciding on an aim of the experiment. They, then, designed their own plan for collecting data and performed the experiment (Task 1). In Task 1, students should design a controlled experiment to find out which insulating material (among three different) is the best to preserve hot water. Pre-service teachers are not provided with detailed instructions by the teacher or the handout (lab guide). This is different from the introductory undergraduate physics laboratory classes in which, after performing an experiment, they should answer some questions. In their physics lab courses, students followed a lab manual with step-by-step instructions and build-in guiding questions to do measurements, calculations, plotting and occasional uncertainty analysis. After the experiment, they prepared and submitted a laboratory report.

The insulation experiment

Investigating how the temperature of hot water falls down in three similar cans each wrapped with three different insulating materials.

1. What variables do you think affect the temperature of hot water in a can wrapped with an insulating material?
2. How does each variable affect the temperature of hot water while it is cooling down and its rate?

I would like you to design an experiment that would allow you to decide which of the three materials is the best for insulation. After you are finished, you need to describe how you plan your experiment, obtain evidence, analyze and explain your data, as well as how you evaluate your experiment.

Task 1. Insulation Experiment

The conducted experiments are those which are usually included in science curricula and mostly they are controlled experiments or simply "fair" tests. The following Table (Table 1) presents a list of experiments.

Table 1. List of Experiments

List of Experiments	
Hooke's law	Exp 1
Free fall	Exp 2
Simple pendulum motion	Exp 3
Insulation experiment	Exp 4
Measurement of resistance - Ohm's law	Exp 5
Refraction and reflection	Exp 6
Friction	Exp 7
Electromagnetic induction	Exp 8

The university is a research public one and the official language is English. Also, teaching and all coursework has been in English.

The reported study is a part of a bigger research project (Gkioka, 2019), which aimed to develop tasks to elicit pre-service physics teachers' understandings of experimental procedure. For the purpose of this paper, the focus will be only on pre-service teachers' evaluation of experimental results. Thus, the participants were taught the criteria of what makes a good quality lab report and about what should be included in each section (i.e., in the evaluation section). Table 2 presents the main points that should be included in the evaluation.

Table 2. The Content of the Evaluation Section of the Laboratory Report

What needs to be included in the evaluation section of the laboratory report?
to look critically at the results to judge how well one can rely upon evidence from the experiment to draw conclusions (E1),
to comment on whether the results are accurate, supporting this with specific data, i.e. % error calculations (E2),
to comment on whether results are valid and reliable-could other factors have affected the results, would this have been significant, is there any evidence for this? (E3),
to identify and account for any anomalous results (due to inaccuracies in the procedure) (E4)
to talk about limitations in the experiment and suggest possible improvements (E5) and finally,
to talk about specific proposals for further work that would improve or add to the evidence (E6).

Research Methods and Data Sources

The study utilized qualitative case study research methods (Stake, 1995; Yin, 2017) in an effort to collect in-depth and comprehensive information about the participants' conceptions of scientific evidence. Stake (1995) defined a case as having specific boundaries in terms of one phenomenon, time and place. For the purpose of this study, the boundaries are the teaching context of the course, within which the research took place. Also, the study is limited to the context of this University and the teacher education system of Turkey. However, in physics education research, undergraduate physics students demonstrate the same difficulties beyond educational systems and countries (McDermott, 2014). The main data sources for this study included:

1. Collection of submitted laboratory reports for their own experiments. The aim was to explore pre-service physics teachers' difficulties in evaluating scientific evidence and writing the evaluation section of laboratory reports. The participants performed some experiments and did the write-up of the laboratory report.

2. Collection of students' work on written tasks in which the participants were asked to judge the quality of data sets and a procedure that other experimenters followed and evaluate those data. The main question was: "Do you trust these data and their conclusions?" It is important for respondents to make their reasoning clear. Such written tasks were developed by the researcher to draw the participants' attention to specific common mistakes and misuderstandings, as documented in the literature.

3. Collection of exam (mid-term and final) papers and finally,

4. When clarification was needed, we collected more data by conducting focused interviews with the participants. Short individual interviews were conducted at different times based on their laboratory reports and coursework. Short task-based interviews have been effective in revealing students' conceptions of scientific evidence. Many of the research studies described in the literature review utilized interview techniques (Allie et al., 1998; Varelas, 1997; Séré, 1999). A growing number of physics education researchers have also recommended the use of task-based interview protocols. This is consistent with prior researchers' support of the use of in-depth interviews that surround the

completion of a task (McDermott, 2014). Thus, such interviews are semi-structured. All interviews were audio-taped and transcribed for analysis. Examples of interview questions are shown in the Appendix.

The principal research investigator was also the instructor of the course. The role of the main researcher and the project assistant was made clear to the participants. The interviews were conducted with informed consent and by following the University Research Ethics Committee protocols. Attention was given to the research ethics (Gregory, 2003) and the associated issues (anonymity of participants and the role of the researcher).

Data Analysis

A qualitative data analysis identified common trends across the participants. The reliability of the analysis is based on the triangulation of the methods employed (Creswell, 1998). Qualitative (Creswell, 1998; Miles & Huberman, 1994) and quantitative data analysis was conducted to identify common themes among the participants. The analysis took place in two different phases. In the first phase, we generated initial categories from laboratory reports, examination item responses and interviews of each pre-service teacher. We constantly compared new data from the laboratory reports and exam items (mid-term and final) with the current categories and refined them. When clarification was needed, we collected more data by conducting focused interviews with the participants. The data were analyzed by comparing the responses for each question both across the interviewees and through each interviewee to identify key categories and features among teachers.

In the second phase, the analysis of laboratory reports and the development of categories were made according to the points that an evaluation section should include (Table 2), according to the author's prior work (Gkioka, 2019). The same frame was also used in the teaching of the course/module.

In the results section, we will present examples of the tasks with which the participants were presented and worked on. We also present excerpts from pre-service teachers' laboratory reports, particularly the section of evaluation. In the excerpts, the participants' language has been kept as submitted (no corrections were made by the researchers in terms of grammar and spelling in their English).

Results

The analysis has revealed a wide range of difficulties among pre-service physics teachers when evaluating secondary sources evidence and writing the evaluation section in laboratory reports. In presenting the main results, the discussion will be around the following headings:

1. Understanding of controlled experiment and related reasoning: Under this heading we have included some misunderstandings related to the same conditions under which the experiment takes place. For example, when students design and plan how to perform the insulation experiment they argued that the top surface of each can should not be covered, because in such way it will be much easier for them to take the thermometer readings without mistakes. Thus, as long as they perform the experiment under the same conditions for all insulated cans and insulating materials, "*there is no problem*" (in student's words). For them, the priority is to ensure that the three cans are under the same conditions and not that the experiment is a controlled one. However, such decision is not consistent with the aim of the experiment, which is to find out which material – among three – is the best for insulation. Such a students' priority restricts them from designing a valid experiment.

The following is an excerpt from an interview with a highly achieving pre-service student:

Interviewer (I): How did you decide to design your experiment?

Pre-service Teacher (PT): I will cover the whole surface of the three cans apart from their top.

I: Why?

PT: There is no problem if you do not cover all the three. You do not need a cover-lid on the top.

I: We are interested in insulation.

PT: But we are interested in the best material. I do not think that they will differ too much, if we cover it or not.

I: Yes, but we are doing insulation.

PT: I do not think think that we should close it.

I: Why? Can you explain your decision to me?

PT: If we put the lid, then taking measurements will be harder, there will be random errors associated with it.

I: The insulating material should be around and at the bottom.

PT: If we do not put the cover on all the three, I think it should be a valid experiment because we do have insulation (because of the covered part) - the top is not necessary.

2. The majority of participants rely upon comparison of prediction with the pattern of the graph to write: "*Our experiment was successful because we found the linear relationship between ..*".

"In all, our experiment was successful by seeing a linear relationship between voltage and current. Anyway, by finding a linear relationship between voltage and current, we can say again that our experiment was successful"

"From our experiment results, we can say that our experiment is successful. By carrying out a fair test, we can trust our data". (Excerpts from their laboratory reports). However, this teacher did not explain why his experiment was a fair test.

In the case of a line graph, there are particular misuderstandings in the evaluation related to what makes a "good" line graph. In particular, students' reasoning is related to whether the line should go through the origin 0 (0, 0). *"We found that our graph has a linear trend. It is passing through the origin".* Many students believe that the fact that the graph passes through the origin is important and gives a definite measure of the success of the experiment. The same reasoning appears in the statement: "*Since the graph does not cross the origin point, I can say that the experiment is not successful".*

The graph may be exponential as in the following evaluation excerpt: *"Our experiment is successful because we get expected result that is our rate of cooling decreases".*

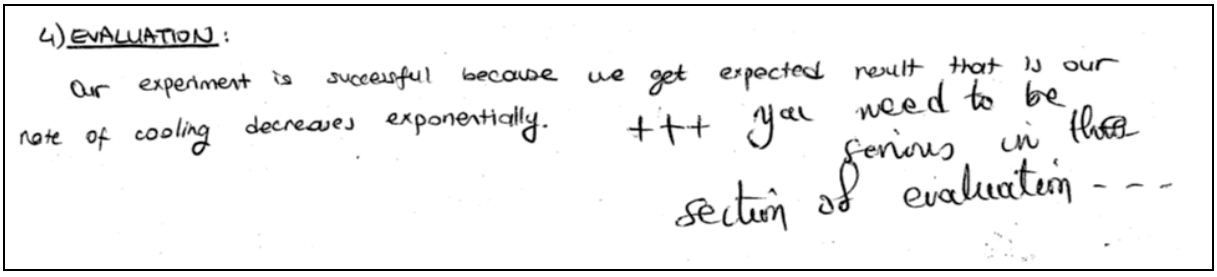


Figure 1. Excerpt from an Insulation Lab Report (Evaluation Section)

Such an evaluation is too short. It is just one sentence in which the pre-service teacher describes the trend of the graph. The instructor's comment is: "*You need to be serious in the evaluation".*

3. Some students rely primarily upon comparisons with theory as the main criterion. They compare their own results with theory. Experimental results are trustworthy when they match with theoretically predicted values. As one participant wrote in the evaluation: "*Results are matching my theory and prediction. It can be said that our experimental results are accurate". "The experimental data match my theory and prediction*". "*I think this experiment is successfully designed to achieve the aim. At the end of the experiment, I figured out that there is a direct proportionality between the current and the voltage. Since my prediction also*

emphasizes that, I can say my experiment is successful in terms of the prediction and the aim". And, *"My data support theory. I got the graph that I would expect. My prediction matches the graph that I obtained. This means that my results are accurate".*

They may mention the term 'errors' without explaining, thus, they conclude: "*Even though I have errors in my experiment, it is easy to see the result that my prediction is correct because there is a linear relation between the length and the squared period of the pendulum".* Initially students talk about "errors" in general, without making the distinction between systematic and random errors. For them, "errors" are like mistakes and wrong measurements, which need to be corrected.

4. Experimental results are trustworthy when similar results are achieved under multiple trials. Students mentioned about repeatability as a main criterion to justify results. "*It is important to have some idea of what the data should look like and that the data is reproducible".* They tried to come to a conclusion by comparing individual measurements between two sets of data, typically reasoning that the "*values for the two groups match almost exactly".* Thus, they do not consider any uncertainty in their measurements. They do not consider both the average value and the range (the overall spread), while they are good at calculating the standard deviation of results. The participants did not consider, refer to the spread, range and uncertainty of results. When comparing data sets, students identify "anomalous" results. They call them wrong measurements, anomalies or mistakes.

5. The pre-service teachers tried to consider errors - systematic and random errors. However, they do not make the distinction between systematic and random errors. They are not successful when they want to argue how errors can be identified and eliminated.

> *"There is always error in any experiment" "… However, human errors are inevitable..." "The anomalous point is a mistake, human mistake. Any experimenter makes mistakes.*
>
> *"Although the graph shows the linear relationship between T^2 and l correctly, I found the g values higher than the theoretical value in all cases. This shows that I have some errors in my experiment. I think my experiment has two types of error which are systematic and random"* (Excerpt from the evaluation of a pendulum lab report).

They think that systematic errors are of bigger values than the random ones: "*Designing the experiment is successful; however, in the experiment, I have some errors. First of all, the resistances of A and B can be acceptable because they match with the theoretical values of the resistances. However, the experimental value and the theoretical value of the C is totally different from each other which is an example of error for this experiment. The theoretical value for C is 56 Ohms; while the experimental one is 70 Ohms. I think such a big difference is not an example of random error. It is a systematic error…"* (from an Ohm's law lab report).

How do they deal with random errrors? Two different ideas appear here: One portion of students argue that random errors can be corrected by taking repeated measurements: "*On the other hand, these anomalous points are examples of random errors. If I repeat the experiment for them, I can exclude the anomalous points…. By repeating the experiment for that points, I am going to exclude the random errors".*

Students many times explain that by repeating they get rid of errors, including systematic ones: "A*lso in my graph, there are two points which are far from my best fit line. These points are the ones that the systematic error is the highest because the g values are between 10.20 and 10.80. To be more accurate, I can repeat the measurement of these data points. Thus, they approach to the best fit line, and my systematic error also decreases".* Here, 45% of the participants are wrong because they have said that by repeating, systematic errors decrease. But systematic errors cannot be eliminated by repeating measurements.

Students often focus on the calculation of errors and miss the big picture of the evaluation. For example, even if they calculate the error, they do not justify their conclusions. Judgements are based on arbitrary criteria:

> *"My percent error is ony 5%, so my experiment proved the theory".*
>
> *"My results are accurate because the error is less than 10%".*

"*None of the graphs goes through the origin of the graph even if they should, according to theory. This is due to random errors such as human errors"*. A highly achieving senior student explained in her evaluation section: "*According to our graph, the best fit line does not pass through the origin but according to the formula, equation T= 2π√l/g, the graph should go through the origin, because of the linear relationship. This is because of errors"*. They are wrong to believe that the line of best fit should go through the origin because there is a linear relationship between the variables of T^2 and length l.

6. They do not distinguish among validity, accuracy and reliability. "*Our experiment, according to our results, is not valid, but our data is accurate and reliable"* (for pendulum motion experiment). The student is wrong here, since s/he needs, firstly, to check whether the experiment is valid. Only then, he needs to examine the accuracy of the experiment by calculating the g values to examine how close the experimental results are to the theoretical value. On this basis, another student is wrong to have said: "*To check the validity of my experiment, I am going to calculate g values from my data sets"*.

When they identify some "different" data, they call them anomalous data. They explain that the data set is not good because they should have yielded the same readings. Therefore, they suggest repetition so that they obtain the same data. Thus, two tendencies appear with regard to how they deal with repeated measurements. One part of students excluded them and calculated the average value which is closer to the "true" value. "*.... I can repeat the experiment for the anomalous points. This will help me to correct the values of them. Maybe anomalous points are the results of misreading the voltmeter or ammeter. I think if I pay attention on these points, my experiment will be improved, so it will be better"*.

The other part, when they identify an anomalous point, they may still include it in the calculation of the average value. In the interview they explained that in the calculation of the average value, all measurements should be included because this is what they had been taught in mathematics classes.

7. Students have had difficulties with accuracy and precision. They do not distinguish between accuracy and precision. For example, a student confused the two notions of accuracy and precision of results, since she argued that: "*It is important to average sets of results because it eliminates inaccuracies. In actually working out an average, we make it more accurate"*. We need to underline here that it should not be taken for granted that any experiment should be repeated. Instead, there should be a real reason that one will repeat an experiment a few more times to get the best average sets of results.

The following is an excerpt from a laboratory report for Hooke's law experiment, which is an improved version, after the instructor gave feedback to the student. Indeed, the pre-service teacher demonstrated a good understanding of accuracy and precision and he is right when he has distinguished between the two concepts of precision and accuracy.

"If I compare the results for my calculations from the table and the graph, my k-values are ranging from 1,4 N/m to 1,6 N/m. The average of the k-values in my experiment is 1,5 N/m. Thus, the results of my experiment show that the constant factor characteristic of spring; i.e. k, is equal to 1,5 ± 0,1 N/m. In other words, I can say that I founded the k-values closer to each other. This shows that my k-values are precise, but I cannot decide whether they are accurate or not. The reason for this is the lack of knowledge about the exact k-value of the spring" (Hooke's law lab report).

Figure 2. Excerpt from Hooke's Law Laboratory Report-Evaluation Section

To explore students' understanding of accuracy and precision, they were asked to work on the following task, which is from a mid-term exam.

Four undergraduate students performed an experiment about Ohm's law and found the value of the resistance. The theoretical value of the resistance is 3.0 Ohm. Four data sets are taken by students given below. Analyze the data sets in terms of their accuracy and precision by explaining why.

Data set 1	Data set 2	Data set 3	Data set 4
2.1 ohm	2.2 ohm	3.0 ohm	3.2 ohm
1.9 ohm	4.9 ohm	3.0 ohm	1.2 ohm
1.7 ohm	2.8 ohm	3.1 ohm	2.4 ohm
2.0 ohm	3.4 ohm	3.0 ohm	3.1 ohm
1.8 ohm	4.1 ohm	2.9 ohm	0.6 ohm

1) Are the measurements accurate and precise? Explain your answer analytically.
2) Tell the four students what they need to do to improve their measurements.

Task 2. Task Related to the Concepts of Precision and Accuracy

In this task, 32 students, out of 36, consider the average value of each data set which they compare with the theoretical value of 3.0 Ohm without considering the range of values. In contrast, only four (4) students looked at the range of each data set.

Table 3. Students' Responses for Task 2

Categories	Number of Students
Consider and calculate average only for all data sets	32
Consider both the average value and the range	4
To repeat for data set 2 and 4	29
To repeat for data set 1	31
To repeat for data set 1 and exclude 1.7 Ohm	5
Only the measurements in data set 3 are accurate and precise	7

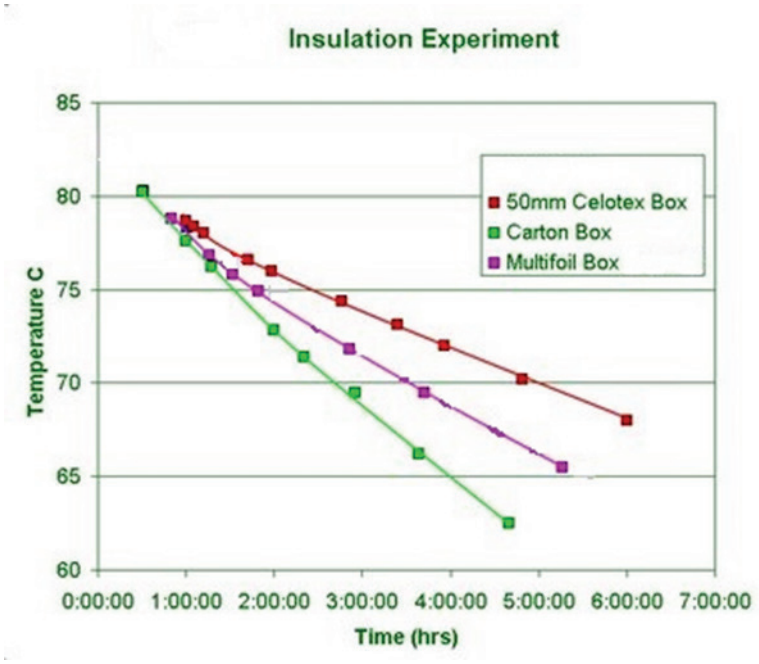
Many times they say that one needs to repeat and calculate the average value but they are not able to reason what it is that "gets better" when one repeats the measurements. It takes much time, from the sixth week, to understand that by repeating one gets more precise results because random errors are eliminated.

All pre-service teachers, including the high achievers, do not distinguish between wrong measurements and uncertainty. Thirty-two (32) did not consider the range of values, only the average value and how close the values are to the theoretical value. Only five students reported that the third data set is good enough, because measurements are close to each other, and close to the theoretical value of resistance.

When in the interview they were asked "If you repeat, how do you process the data?". They replied that, if they have sets of repeated measurements, they calculate the average value and then, they plot the average values to draw the best bit line. But in this way, the results are "distorted"; instead, they should have plotted all the points. Then, they will be able to see more clearly the trend or the pattern of the data.

One more task (from the final exam) was designed to explore how students evaluate experimental evidence and procedure. With this task we wanted to investigate whether students are able to write what the aim of the experiment is, by using physics theory about insulation. Then, whether they can evaluate the design of the experiment and the quality of evidence and judge the accuracy of the experiment, analyze data and explain data by using theory in order to reach the aim of the experiment to determine which of the three given insulating materials is the best. Further, in this task (Task 3), pre-service teachers are given secondary sources of evidence and they are asked to evaluate the whole procedure and finally, make suggestions for improvement of the experimental procedure.

Three bottles of tea at the same initial temperature are completely wrapped with three different insulating materials of one layer.
By using a long thermometer (put in each bottle through a little hole), temperature data is taken for each bottle for six hours. The measurements taken have been plotted and the graphs are shown in the figure below.
1) Do you trust the data?
2) If you carry out this experiment by yourself, how will you improve it?



Task 3. Insulation Experiment - Use of Secondary Sources Experimental Evidence

The main finding is firstly the misunderstanding that accuracy is closely related to the number of taken readings. In their own words: "*The larger the number of readings, the greater the accuracy of the time achieved for the accuracy of the experiment*". *And, "The more measurements you take the more you know how accurate you are". "Repeated measurements improve the accuracy of results"*, which are not correct. They also argued that more measurements are needed. In the interview, when they were asked to specify how many and explain, thirty-three (33) of them said: "*As many as you can*".

Table 4. Pre-Service Teachers' Suggestions for Improvement of the Experiment

Suggestions for Improvement	Number of Students
Take more measurements	36
Take measurements more often	32
Take measurements for longer time	36
Take as many measurements as you can	33

While making the above suggestions that more and more regularly taken measurements need to be taken and for longer, they have not explained what will become better. In other words, how the quality of data will be improved. However, in this task, they should have explained that one needs to repeat the experiment at the same room temperature to reduce systematic errors and then, perform the experiment several times to reduce human errors.

8. For all students, with no exception, the priority in evaluations is to talk about mistakes in the experimental procedure they followed. For example, in the insulation experiment, they were touching the thermometer, and they should not have to. Also, that the temperature readings should have had one

more significant figure and so on. But they did not make any judgement about to what extent these affect the quality of the collected evidence and their conclusions. In their written lab reports, students discuss the limitations of the data they collected and the variables that they may have affected results (such as room temperature variation over 24 hours) but explained that those were difficult to control or not considered. Thus, they want to suggest improvements in the followed procedure and extension work (i.e. look at another variable). However, here, apart from talking about the difficulties or the problems they possibly encountered, they should explain how such difficulties affected their results.

Generally the tendency towards evaluating is to find things that are wrong with it and suggest what they should change. Students have to suggest improvements to the method followed and should also propose extension work. They list and explain anomalous results and suggest improvements on the method and extension work. Students also have to talk about any difficulties they encountered. However, the findings from this study show that students do not get a very clear understanding that even with experiments where you know certain things could be improved, you can still make the point that those factors are not significant. That is, overall, they have not affected the validity of the results. By improving those factors (i.e. room temperature to be kept constant) one would nearly get slightly more accurate results. In addition, the fact that the temperature has not been strictly controlled during the experiment does not appear to have affected the reliability of the results. Apart from suggestions for improvement in method and extension work, the majority of students did not consider the reliability and validity of results as the first priority. Even when a few of them did (four students in Ohm's law and insulation experiments), they did not explain how such reliability and validity support their conclusions. As they explained in the interviews, they found it difficult to put all of the points (E.1 - E.6) together and articulate their ideas within the evaluation section.

However, evaluating experimental evidence concerns primarily the validity and reliability of results to support the drawn conclusions. If we restrict evaluation to describing features of the data or experiment (imperfections or variability of data) or suggesting how to improve an experiment, as pre-service teachers do, then the question follows: "If this is the way to improve, why did not you try to improve it from the first time?". In other words, suggesting improvements for an experiment they have just carried out does not make sense.

The above results can be presented in the following Table (5) which shows the distribution of evaluation categories (E.1 - E.6), as presented in Table 2 for the evaluation content for eight (8) experiments (see Table 1) for all pre-service teachers (N=36).

Table 5. Distribution of Evaluation Categories (E.1 - E.6) in Laboratory Reports for Eight (8) Experiments for All Pre-Service Teachers (N=36)

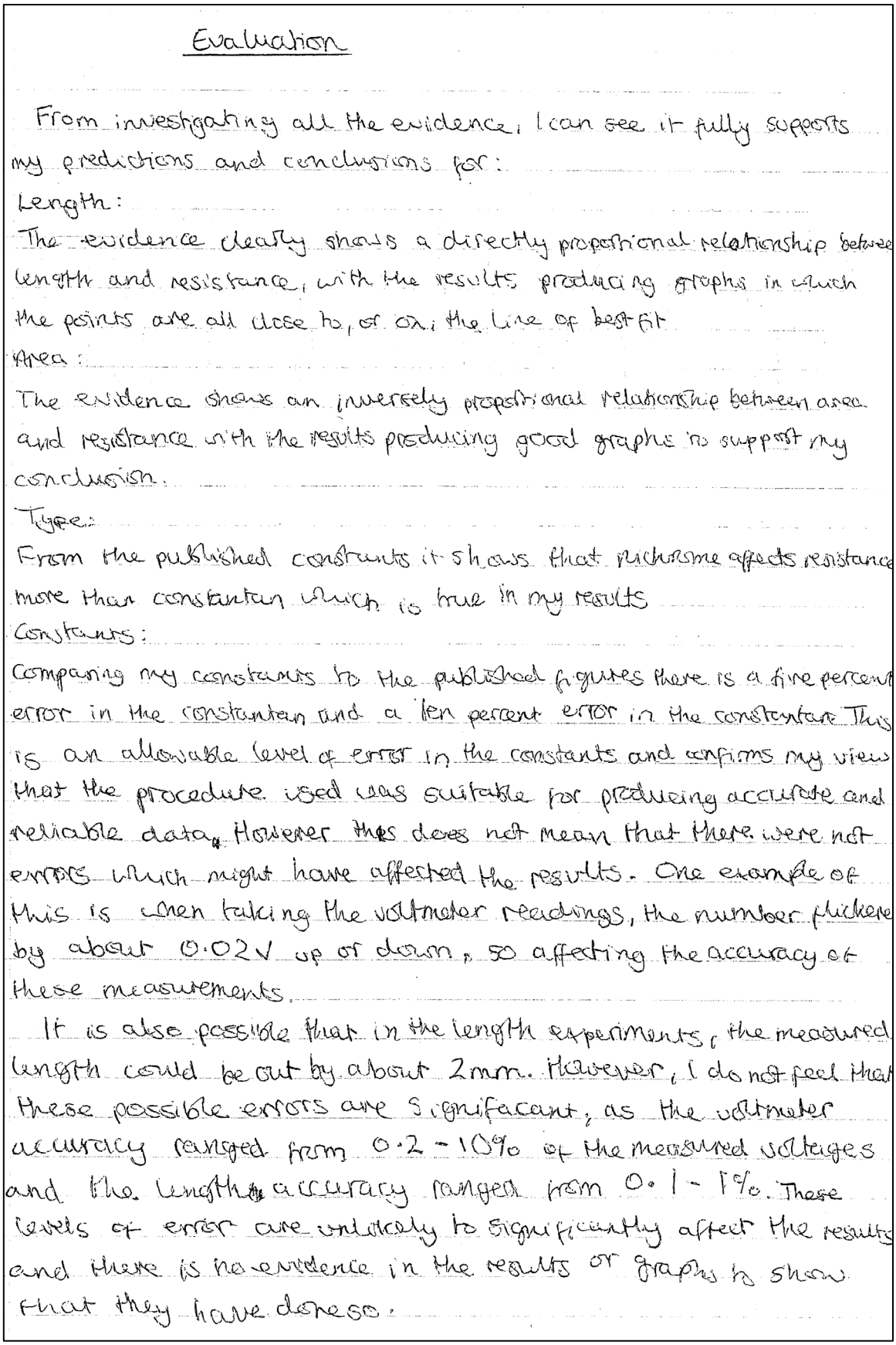
What an evaluation section needs to include		**Number of Students**							
		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7	Exp 8
To look critically at the results to judge how well one can rely upon evidence from the experiment to draw conclusions.	E.1	2	2	2	2	2	2	2	3
To comment on whether the results are accurate, supporting this with specific data, e.g., % error calculations	E.2	3	2	3	3	4	4	2	2

Table 5. Continued

What an evaluation section needs to include		Number of Students							
		Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7	Exp 8
To comment on whether results are valid and reliable - could other factors have affected the results, would this have been significant, is there any evidence for this?	E.3	2	2	3	2	4	3	2	3
To identify and account for any anomalous results (due to inaccuracies in the procedure and/or measurements).	E.4	34	34	34	34	34	34	34	33
To talk about limitations in the experiment and suggest possible improvements (specific detail required).	E.5	36	36	36	36	36	36	36	36
To talk about specific proposals for further work that would improve or add to the evidence.	E.6	32	31	31	32	32	32	32	29

The evaluation categories E.1, E.2 and E.3 are usually neglected by the participants. They talk about errors and sources of errors (E.4), as well as they focus on and they talk extensively about limitations in the experiment and suggest possible improvements (E.5).

In concluding this section of results, I would like to present an excerpt of a very good piece of evaluation (from an Ohm's law report), in which the pre-service teachers has included all the points from E.1 to E.6.



Evaluation

From investigating all the evidence, I can see it fully supports my predictions and conclusions for:

Length:

The evidence clearly shows a directly proportional relationship between length and resistance, with the results producing graphs in which the points are all close to, or on, the line of best fit

Area:

The evidence shows an inversely proportional relationship between area and resistance with the results producing good graphs to support my conclusion.

Type:

From the published constants it shows that nichrome affects resistance more than constantan which is true in my results

Constants:

Comparing my constants to the published figures there is a five percent error in the constantan and a ten percent error in the constantan This is an allowable level of error in the constants and confirms my view that the procedure used was suitable for producing accurate and reliable data. However this does not mean that there were not errors which might have affected the results. One example of this is when taking the voltmeter readings, the number flickered by about 0.02V up or down, so affecting the accuracy of these measurements.

It is also possible that in the length experiments, the measured length could be out by about 2mm. However, I do not feel that these possible errors are significant, as the voltmeter accuracy ranged from 0.2 - 10% of the measured voltages and the length accuracy ranged from 0.1 - 1%. These levels of error are unlikely to significantly affect the results and there is no evidence in the results or graphs to show that they have done so.

Figure 3. Excerpt from a Very Good Ohm's Law Lab Report – Evaluation Section

Discussion and Conclusion

The findings from this study support the idea that pre-service physics teachers experience a range of difficulties when evaluating experimental evidence. Such results are consistent with findings from other studies into students' inability to critically evaluate scientific claims (Albers et al., 2008; Hu & Zwickl, 2018; Leach, 1999). For example, we confirm findings by Hu and Zwickl (2018) who identified several differences in students' ideas abut validity and uncertainty in measurements. The majority of their introductory students justified the validity of results through agreement with theory.

When evaluating experimental evidence, the main point is to look critically at the validity and the reliability of the evidence. That is, one has to look for how well one can rely upon the evidence from the experiment to draw valid conclusions. This means that in the evaluation, it does not have to be automatically assumed that there are improvements that can be made or that worthwhile suggestions for extension work can be made. Nor should it be automatically assumed that there are anomalous results without making reference to the graph.

Students should explain why they believe that their results are valid and reliable. For this purpose, they should comment on whether results are reliable and valid and hence, they have to consider other factors that could have affected the results. Yet, they have to judge whether there is any evidence that such factors would have been significant or not to the validity and reliability of results. For instance: "I acknowledge that the temperature was not controlled, but I do not feel that it had a significant influence on … I would have said that what my results show are valid and that this is the real tricky one, the experimenter to think about which variable s/he had not perfectly controlled, but it does not prevent him/her making a valid conclusion, which means to find something wrong with it".

Our participants have undervalued the importance of particular points: "*How do you know that your results are valid and reliable?" "Is there any evidence that other factors could have affected your results?" "Could such factors have been significant with regard to the validity and reliability of the results?"* Such questions and issues are the key to a good evaluation. Students and teachers should also be aware that even with experiments where there are certain things that could be improved, one might still make the point that such factors may not have been of significance if they had not affected the overall validity of the results. Thus, by improving such factors one may get slightly more accurate results rather than improved validity.

We want to argue that the first issue to be addressed in the evaluations is the validity and reliability of evidence, which, in essence, have to do with the criteria that one uses to determine validity: "*How do you know that your results are valid and reliable?"*. A good level of understanding of experimental validity and measurement reliability is the foundation for good quality evaluations in laboratory reports (E.1 - E.3 points). Suggestions about improvements and discussion of limitations of the experiment have to come later (E.4 - E. 6).

However, the majority of our participants did not take a critical look at the results to make judgements about how valid and reliable their conclusions were (E.1). One has to consider that data allows him/her to draw a valid conclusion. That the method used is appropriate but certain small changes in accuracy could be made. This is really a neglected aspect of their evaluation section. Further, they did not comment on whether results are valid and reliable - could other factors have affected the results, would this have been significant, is there any evidence for this? (E.3). However, all of them made suggestions for improvements and talked about specific proposals for further work that would improve or add to the evidence (E.4 and E.5). Our participants may have identified anomalous results and explained them, they may also have suggested improvements in method and extension work, but they still find it difficult to put all of them together and articulate their ideas within the evaluation section. In a few words, our participants have lost the whole picture of the evaluation.

Most students can obtain a few marks for anything valid they have written about the experiment and the followed procedure. But, in order to obtain a higher mark they have to examine and justify

whether the results are valid and reliable. Instead of students' talking only about improvements in the followed procedure of the experiment, they should explain why they did not make improvements from the beginning. Instead of writing: "I'll improve that and so on, I will get marks, so temperature was not controlled and it should have been", they should provide clear evidence that they are actually doing it initially and saying: "Temperature was not controlled but the results suggest this was not important factor, because any changes would have no significance".

As teachers and physics educators, we should not be mistaken that our students are competent at evaluating experimental evidence because they were already taught about such issues in the undergraduate physics lab classes. Nor that have they understood the reasoning behind average calculations and methods of best fit plotting in their mathematics classes. We do not want to argue that undergraduate physics laboratory or mathematics classes are not effective. Instead, we want to make an argument about the need for the two Departments of Physics and Physics Teaching to work together towards the development of experiments and tasks in which students will have more flexibility and responsibility for designing the plan, conducting experiments, analyzing results and evaluating evidence. In experiments, physics students should make decisions on strategies for establishing validity, such as repeatability and uncertainty calculation. Decisions in the planning - which variables to keep constant and which to change - have also to be explained in the evaluation with regard to the validity of the experiment. In the teaching practice we should give our students the possibility to make choices to design their own experiments so that they address issues of validity and reliability. A similar argument was developed by Tiberghien, Vieillard, Le Marechal, Buty, and Millar (2001) who undertook a survey of laboratory tasks in senior secondary school and undergraduate courses across Europe. They found out that less than 15% of experimental tasks in physics laboratory courses require students to design their own experiment (Tiberghien et al., 2001). These findings call for the development of new teaching materials and tasks.

In addition to discussing the experiments, we want to underline that we need to modify the traditional curriculum. Students need practice in evaluating their own results or secondary sources evidence. The process requires time. Inevitably, this constraint places a limit on both the breadth of material that can be covered and the pace at which instruction can progress. Tasks need to emphasize the ability of reasoning in order to improve the match between teaching and learning. As teachers, we are faced with the challenge to teach less but in more depth.

These results, taken together, suggest that the present form of the laboratory courses was not able to develop good evaluation skills in pre-service physics teachers.

Teaching practice should focus on a clear structure in the evaluation section and further, on the need for drafting and re-drafting it. In fact, instruction focused on how to write evaluations needs to be a long-term process which has to begin with much direction and help on the part of the teacher. Later on and by time, more freedom and flexibility should be given to students. As Kalthoff et al. (2018) have argued, various more or less explicit instructional approaches need to be introduced for the promotion of experimental skills in the training of school teacher students.

Only two students (out of thirty-six) made a judgement about the validity and the reliability of the results in their experiments. They clearly evaluated the methods used to collect results, for example, recognized that temperature environment is not constant, but changes over twenty-four (24) hours and explained that this does not affect the quality of results. The same students in their written laboratory reports looked critically at the results to judge how well they could rely upon evidence from the experiment to draw conclusions. Toward such direction, they made calculation of errors to support claims about accuracy of results. Thus, such points seemed to be the most difficult parts of the evaluation section.

Perhaps the most serious difficulty that we have identified is failure to integrate related concepts into a coherent framework to form the evaluation section in the laboratory report. Pre-service physics teachers have not been able to make the distinction between validity and reliability, as well as

between accuracy and precision. In addition, many times they do not understand "why" they do what they do. For example, why they need to take repeated measurements. They seem that they use rules by rote, as well as mehods of processing results and plotting points to draw a best fit line. The students rely on rules that they have incorrectly memorized. For example, that the best fit line should go through the origin. The analysis of students' responses to the presented tasks gave us several insights into students' views about the criteria for establishing trustworthy results. Their difficulties also concern a failure to make the distinction between random and systematic errors. Also, their understanding of uncertainty was limited and consistent with what Caussarieu and Tiberghien (2017) reported when they explored how a first-year university physics course deals with measurement uncertainties. Uncertainties were often missing from the results of students' calculations.

Similar findings were reported by Séré, Journeaux, and Winther (1998) who argued that the fact that students apply the calculation aspects of data analysis does not imply that they fully understand the underlying principles behind meaurement reliability.

The present discussion suggests that we need to plan different experiments and tasks in which there will be a real reason for repeating the experiment, plotting the points and working out the average. A suitable experiment to teach the concept of average is one where the plotted readings lay either side of the line to be drawn. This strategy may lead students to say that the quality of their results is not good and thus the teacher can get students to suggest to repeat the experiment in order to eliminate random errors. Thus, an irregular pattern of a line graph can be a good prompt for teaching the need for repeating the experiment (i.e. Hooke's law or Ohm's law) a few times. Then, they are introduced in averaging techniques. Students should also be explicitly taught about how to work out averages without assuming that averaging methods were grasped in previous maths lessons. The last step is to produce a best-fit line to obtain a visual pattern.

One step further is to be able to appreciate that by taking repeated measurements and taking the average helps eliminate random errors but not necessarily inaccuracies in the method or faults that affect the reliability of the results. Students and teachers often misunderstand the concepts of accuracy and reliability, believing, for example, that repeating readings make them more accurate and more reliable, whereas, all it does is help to check repeatability. This is because they appear to be arguing that repeating results three times is a method of improving reliability which may, of course, help identify individual errors of the experiments, where an error was made, but it does not automatically improve the reliability. When there are large variations, it is necessary to review and modify the method.

One can improve precision and get closer to the 'true' value by repeating many times and taking a mean average value. By repeating, one is increasing the likelihood that she will take measurements at or near the 'true' value. Then, the average that one gets will be closer to this value.

Reliable readings have to encompass all the ideas of controlled experiment, accuracy and precision. A reliable set of readings gives close to the same values when repeated. However, results that are repeatable are not necessarily reliable: they could all be consistently wrong if, for example, a meter was not zeroed properly. The readings must be both accurate and reliable.

Secondly, this study points to the value of developing and introducing tasks which address the concepts of reliability and validity in different contexts in order to teach them about evaluating experimental results in different experiments. We should help our students see common issues in their reasoning they are using in different experiments. Such an endeavour requires time for pre-service teachers to reflect upon these experiences and be able to generalize from them. As Holmes and Wieman (2018) argued, teachers should give time so that students reflect on their results and develop meaningful understandings of what they have been doing and why.

However, most laboratory curricula emphasize the development of laboratory procedures with little attention being paid to develop an understanding of the deeper reasons for these procedures. Such

laboratory courses tend to emphasize the formalistic rules of the statistical procedures of data and omit to include aspects that address a deep reasoning underlying the experimental procedures. Therefore, laboratory curricula need to be designed in which the underlying experimental procedures are explicitly addressed.

Although most pre-service teachers have taken undergraduate courses with science lab classes, these all too often follow a strictly prescribed procedure-which they follow without understanding what they do and why (Tobin, 1990) and without thinking.

As Eshach and Kukliansky (2018) argued, understanding of their difficulties when evaluating, may be of great help to educators in designing better learning environments to develop related laboratory skills. The teacher should be aware of the problems students experience in writing evaluations. There is the need to revise the syllabus of the course on the basis of pre-service teachers' needs and difficulties. We need to explore and study such understandings so that we use them as a starting point for teaching. In the same way that physics education research has investigated "misconceptions" and reasoning in various areas of physics topics.

The write-up of evaluations is a demanding task. It requires much time and effort on the part of students and the teacher. Teaching should give importance to the difficult part of evaluation for students, in particular to the notions of accuracy, reliability and validity and how these should be connected and inter-related to provide a meaningful evaluation section. However, since the evaluation is the final part of the investigation, students may not have enough time and hence, they do not give appropriate attention. Even teachers may run out of time in their teaching. Experiments are usually time-consuming and there is no time left to discuss possible sources of errors.

Lastly, this research wants to argue that when teaching how to deal with anomalous points, the term of "outlier" (rather than 'anomalous measurement') seems to help more because the outlier is defined in relation to a range of measurements. Therefore, they need to define an acceptable range of measurements. An outlier is defined as a value in a set of data that is judged to be usually large or unusually small in comparison with most of the other values, for whatever reason. In contrast, an anomalous value is a measured value that appears not to fit the pattern of the other measurements, and is often (though not always) due to a mistake. For example, a value that is very different from the others in a set of repeated measurements, or a data point that lies far from a line of best fit.

As a whole, the findings lead to the same thought by McDermott (1991) that the difference between what is taught and what is learned is often greater than most instructors realize:

> *"… what the instructor says or implies and what the student interprets or infers as having been said or implied are not the same … There are often significant differences between what the instructor thinks the students have learned in a physics course and what students may have actually learned*" (McDermott, 1991, p. 303).

Teachers usually take for granted that students know how to evaluate evidence. However, in the introductory labs performed by students in first- and second-year laboratory courses, students had little or no autonomy in designing and performing an experiment. They do not write the full lab report, but they answer some questions. For example, after the simple pendulum motion, the questions ask them to make calculations of period T and constant of acceleration and then, to calculate the average of g and the % deviation for g. Then, they discuss and write down possible sources of systematic error. Or, after performing the resistance experiment, the participants are asked to make calculations and then, calculate the average R and the standard deviation of the resistance value.

Students need to develop a solid understanding of sources of errors and the ways they can be eliminated. The effects of systematic errors cannot be reduced by repeating measurements. Instruments should be checked for errors before they are used. If the systematic errors are due to the procedure, revision of the method should be made.

Our conclusion is that this complex topic requires further attention. It seems that a larger investment of time than the one supplied within the framework of our study is called for. For example, there may be the need for further research to follow the participants when they obtain a teaching post in a school or in a college.

Evaluation of experimental procedure is an important part of laboratory work. Writing of the evaluation section should deserve particular attention by physics educators. If we want students to develop meaningful understandings of the evaluation, we should engage them in experiments where they apply this procedure, but we also need to help them think about and examine the reasoning behind this procedure in a way that connects experimental situation and the evaluation behind it with the overall understanding (E. 1- E.6). There should be courses so that preservice teachers have more opportunities to develop evaluation and experimental skills. Like Hofstein and Lunetta (2004), we believe that: "*The literature has suggested that inconsistencies between teachers' goals and behaviours and limitations in teachers' skills, in this case in the school laboratory, should be addressed carefully in long-term professional development programs designed to develop the understanding, knowledge and skill of professional teachers"* (Hofstein & Lunetta, 2004, p. 45).

This study may contribute to those who wish to design a laboratory course for physics laboratory data analysis addressed to teacher preparation. Curriculum developers should be aware of the difficulties students have with evaluation. We need not only to revise the course but develop one research-based module or more for initial teacher education of physics teachers. Pre-service physics teachers need to be taught about secondary students' difficulties in evaluating experimental evidence. This study should inform a new course and vice versa; students' achievement and teachers' experience should inform new research.

Acknowledgements

The study was carried out as a start-up project funded by BAP (Boğaziçi Araştırma Proje) 10800.

I am grateful to the anonymous reviewers for their critical reading of the manuscript. I am also indebted to the preservice teachers with whom I worked, for all their time and effort, without which this study could not have been done.

References

Albers, C., Rollnick, M., & Lubben, L. (2008). First year university students' understanding of validity in designing a physics experiment. *African Journal of Research in Mathematics, Science and Technology Education, 12*(1), 33-54.

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

American Association of Physics Teachers. (2014). *AAPT physics education report: Recommendations for the undergraduate physics laboratory curriculum*. American Association of Science Teachers: College Park, MD.

American Association of Physics Teachers. (2017). *Physics and 21st century science standards.* American Association of Science Teachers: College Park, MD.

Bevington, P. R., & Robinson, D. K. (2003). *Data reduction and error analysis for the physical sciences* (3rd ed.). Boston: McGrawHill.

Boudreaux, A., Shaffer, P. S., Heron, P. R. L., & McDermott, L. C. (2008). Student understanding of control variables: Deciding whether or not a variable influences the behavior of a system. *American Journal of Physics, 76*(2), 163-170.

Caussarieu, A., & Tiberghien, A. (2017). When and why are the values of physical quantities expressed with uncertainties? A case study of a physics undergraduate laboratory course. *International Journal of Science and Mathematics Education, 15*, 997-1015.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* London: Sage Publications.

Crujeiras-Pérez, B., & Jiménez-Aleixandre, M. P. (2017). Students' progression in monitoring anomalous results obtained in inquiry-based laboratory tasks. *Research in Science Education*, 1-22.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school. Learning and teaching science in Grades K-18.* Washington, DC: National Research Council.

Eshach, H., & Kukliansky, I. (2018). University physics and engineering students' use of intuitive rules, experience, and experimental errors and uncertainties. *International Journal of Science and Mathematics Education, 16*, 817-834.

Foulds, K., Gott, R., & Feasey, R. (1992). *Investigative work in science - a report by the exploration of science team to the national curriculum council.* Durham: University of Durham.

Gatsby, L. (2012). *Science for the workplace*. London: Gatsby Charitable Foundation.

Gkioka, O. (2019). Preparing pre-service secondary physics teachers to teach in the physics laboratory: Results from a three-year research project. *AIP Conference Proceedings, 2075*(1), 180009. doi:10.1063/1.5091406

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

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

Grant, L. (2011). *Lab skills of new undergraduates: Report on the findings of a small scale study exploring university staff perceptions of lab skills of new undergraduates at Russell Group Universities in England.* London: Gatsby Charitable Foundation.

Gregory, I. (2003). *Ethics in research*. London: Continuum.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28-54.

Holmes, N., & Wieman, C. E. (2018). Introductory physics labs: We can do better. *Physics Today, 71*(1), 38-45.

Hu, D., & Zwickl, B. M. (2018). Examining students' views about validity of experiments: From introducory to Ph.D. students. *Physical Review Physics Education Research, 14*, 010121.

Kalthoff, B., Theyssen, H., & Schreiber, N. (2018). Explicit promotion of experimental skills. And what about the content-related skills? *International Journal of Science Education, 40*(11), 1305-1326.

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

Lewin, W., & Goldstein, W. (2012). *For the love of physics*. New York: Free Press.

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

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

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

McDermott, L. C. (1990). A perspective on teacher preparation in physics and other sciences: The need for special science courses for teachers. *American Journal of Physics, 58*, 734-742.

McDermott, L. C. (1991). Millican Lecture 1990: What we teach and what is learned-closing gap. *American Journal of Physics, 59*, 301-315.

McDermott, L. C. (2014). Melba Newell Phillips Medal lecture 2013: Discipline-based education research - a view from physics. *American Journal of Physics, 82*(8), 729-741.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* London: SAGE.

National Research Council. (2001). *Educating teachers of science, mathematics, and technology: New practices for the new millenium.* Washington, DC: National Academy Press.

Priemer, B., & Hellwig, J. (2018). Learning about measurement uncertainties in secondary education: A model of the subject matter. *International Journal of Science and Mathematics Education, 16*, 45-68.

Roberts, R., & Johnson, J. (2015). Understanding the quality of data: A concept map for 'the thinking behind the doing' in scientific practice. *The Curriculum Journal, 26*(3), 345-369.

Séré, M-G. (1999). Learning science in the laboratory: Issues raised by the European Project "labwork in science education". In M. Bandiera, S. Caravita, & M. Vicentini (Eds.), *Research in Science Education in Europe* (pp. 165-174). Dordrecht, the Netherlands: Kluwer Academic Publishers.

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

Séré, M-G., Journeaux, R., & Winther, J. (1998). Enquête sur la pratique des enseignants de lycée dans le domain des incertitudes. *Bulletin de l'Union des Physiciens, 92*, 247-254.

Singer, S. R., Hilton, M. L., & Schweingruber, H. A. (Eds.). (2005). *America's lab report: Investigations in high school science.* Washington, DC: National Research Council.

Stake, R. E. (1995). *The art of case study research.* London: Sage.

Taylor, J. R. (1997). *An introduction to error analysis* (2nd ed.). Sausalito, CA: University Science Books.

Tiberghien, A., Vieillard, L., Le Marechal, J-F., Buty, C., & Millar, R. (2001). An analysis of labwork tasks used in science teaching at upper secondary school and university levels in several European countries. *Science Education, 85*, 483-508.

Tobin, K. (1990). Research on science laboratory activities: In pursuit of better questions and answers to improve learning. *School Science and Mathematics, 90*, 403-418.

Varelas, M. (1997). Third and fourth graders' conceptions of repeated trials and best representatives in science experiments. *Journal of Research in Science Teaching, 9*, 853-872.

Yin, R. K. (2017). *Case study research and applications: Design and methods*. Los Angeles: Sage.

Appendix. Interview Questions

Do you trust these data and their conclusions?
What sort of difficulties do you experience when you design your experiment without getting guidelines?
What do you think should be included in the evaluation section of a laboratory report?
How can you make sure that your results are accurate?
Are there any more variables which affect your results but you did not consider them in your design?
Why do you repeat measurements?
If you repeat, how do you process data?
What is it that “gets better” when you obtain repeated measurements?
How do you identify errors?
How can you eliminate random errors?
How do you eliminate systematic errors?
What does 5% error mean to you?
What does 22 % error mean to you?
What do you mean when you write “My results are accurate and precise”?
What would you recommend these students need to do?
What makes the evaluation of an experiment hard?
What kind of help would you need from the instructor?