Science Education International
36(3)_2025,349-360
https://doi.org/10.33828/sei.v36.i3.10

ORIGINAL ARTICLE

Why and Where Future Teachers Fail to Successfully Design Experiments of Complex or Counterintuitive Everyday Life Problems

Ioannis Lefkos*

Department of Educational and Social Policy, University of Macedonia, Thessaloniki, Greece

***Corresponding Author:** lefkos@uom.edu.gr

ABSTRACT

This study investigates the ability of future primary school teachers to design experiments when confronted with everyday life problems, particularly those involving complex phenomena or counterintuitive outcomes. Data were collected using a content analysis approach from the written assignments submitted during a Science Education university course, where students were asked to formulate hypotheses and design relevant experiments based on worksheets having a specific structure. Their responses were assessed using a six-dimensional rubric that evaluates key elements of experimental design, including hypothesis formulation and variable manipulation. Findings indicate that while students generally performed well across most dimensions, they encountered significant difficulties in two: forming hypotheses and identifying/manipulating variables. Non-parametric statistical analysis revealed that these challenges are strongly linked to the nature of the problems. Specifically, hypothesis formulation was hindered in problems with counterintuitive outcomes, whereas variable manipulation became problematic in tasks with higher complexity, i.e., involving more variables. These results demonstrate that these failures are not merely a procedural deficit, but are rooted in the interplay between conceptual knowledge and metacognitive shortcomings in handling complexity. They also underscore the need for targeted interventions and pedagogical reform in teacher education, focused on developing higher-order experimental competence, which can help future teachers become more comfortable with designing tasks, hopefully integrating them in their future classrooms.

KEY WORDS: Assessment rubric, experimental design, future-teachers training, inquiry-based learning, variable manipulation

INTRODUCTION

Although inquiry-based approaches are considered appropriate for promoting not only conceptual understanding but also experimental skills of students (Sokołowska, 2018), the educational community does not seem willing to adopt them (Melville et al., 2013), presenting as a primary obstacle the limitations of teaching time (Ha and Kim, 2020). Altering these perspectives can be greatly assisted by good practices during the education of prospective teachers at the university level, providing them with more chances to enhance their familiarity with inquiry-based methodologies, training them in designing inquiry-based approaches (Molohidis and Hatzikraniotis, 2018), or in designing appropriate assessment tools depending on the learning objectives they set (Harlen, 2013).

However, the importance of the inquiry-based approach is not limited to future scientists (students or pupils), but is addressed to everyone, as it is an approach that can be applied wherever problems need to be solved, whether in everyday life (Abd-El-Khalick et al., 2004) or in professional life (Karelina and Etkina, 2007).

For the education of future primary school teachers (FPTs), competence in science is particularly important, especially in mastering basic science process skills (SPSs) (García-Carmona et al., 2024). Teachers who lack proper preparation in science might be limited in their ability to create quality learning environments and may unintentionally pass on inadequate or incorrect scientific concepts to their students (García-Carmona et al., 2024).

Inquiry-based Learning (IBL) and Laboratory work

Curriculum reforms worldwide are increasingly promoting IBL (Agustian et al., 2022; Girault et al., 2012; Hatzikraniotis et al., 2010), responding to the widespread criticism of traditional laboratory instruction, which often relies on "cookbook" approaches emphasising following prescribed protocols rather than authentic scientific practice (Agustian et al., 2022; Bitzenbauer and Meyn, 2021). IBL seeks to promote deeper, more meaningful learning by involving students in scientific practices and encouraging a transition from teacher-led instruction to student-driven investigation (Agustian et al., 2022; Girault et al., 2012; van Riesen et al., 2018a).

For IBL, laboratory work and experimentation are an essential part, as they are considered a very important factor in learning (Efstathiou et al., 2018). This is because not only do they bring learners closer to the scientific content, but they also help them understand the nature of science (Millar,

2004). Science education researchers commonly agree that a thorough understanding of science involves not only learning concepts and models ("knowing science") but also actively "doing science" and "knowing about science" (Psillos, 2023). Therefore, laboratory work is recognized as an essential part of physics (Psillos, 2023; Cai et al., 2021), chemistry (Agustian et al., 2022), and biology education (Brownell et al., 2014; Dasgupta et al., 2014), promoting student competencies, like experimental abilities and higher-order thinking skills (Agustian et al., 2022).

The Role of DOE in Laboratory Work

Several models have been proposed to describe or model the experimental process (an interesting analysis has been done by Emden and Sumfleth (2016). These models might have several dissimilarities, but in most of them, the design phase of the experiment (plan or design) is included. In addition, in some models, the planning phase includes the formulation of a hypothesis (or question), while in others, the hypothesis is regarded as a distinct phase. Furthermore, the execution phase is also integrated into the design of the experiment in certain models. In this paper, we consider formulating a hypothesis to be included in the design phase of the experiment. On the other hand, we consider the execution of the experiment to be a separate phase, following similar proposed models (Kipnis and Hofstein, 2008).

Design of Experiments (DoE) can be seen as a distinct phase (Lefkos, 2024) of conducting scientific investigations (Agustian et al., 2022), is recognized as a key scientific skill (Brownell et al., 2014; Dasgupta et al., 2014) and essential for developing higher-order cognitive skills (Girault et al., 2012; Koretsky et al., 2008; Shanks et al, 2017). As an integral part of scientific inquiry, the DoE includes various phases, such as generating hypotheses, selecting and controlling variables, choosing appropriate methods and tools, and drawing conclusions (Aydoğdu, 2015; Dasgupta et al., 2014; Deane et al., 2014; Koretsky et al., 2008; Lefkos et al., 2011). Providing DoE training to future teachers is especially beneficial, as it can encourage broader adoption of IBL in future educational settings (Lefkos, 2024).

DoE as an Autonomous Activity

Viewing design as an autonomous stage facilitates the swift and straightforward execution of relevant activities, not dependent on the execution of the experiment, and thus more accessible to the training of appropriate skills within the often tight time frame of the teaching process (Ha and Kim, 2020). Furthermore, independent design activities can be implemented without regard for the laboratory modality (i.e., physical or virtual) in which the subsequent experiments will be conducted or the methodological approach to experimentation (i.e., explicit or implicit) (Karagianni and Psillos, 2022).

The design of an experiment for a given problem has been proposed as a higher-order skill than the actual execution of an experiment (Garratt & Tomlinson, 2001). This phase of the inquiry cycle is arguably the most challenging (Efstathiou et al., 2018) and, therefore, one of the most essential elements of inquiry (van Riesen et al., 2018b). The design process is an advanced cognitive endeavour triggering critical thinking skills, significantly different from merely following instructions, as seen in traditional laboratory practices, which primarily focus on mastering laboratory techniques in conjunction with instrument functionality (Komives, 2015). The demand for critical thinking and reflection in the DoE leads some researchers to define it as an "ability" (Etkina et al., 2006), indicating that it is more than a skill. This view is also one that we adopt in this paper.

The design of an experiment to solve a given problem involves several individual phases, as follows: (a) students first have to formulate a hypothesis, (b) identify the variables involved, and apply a strategy to control them, thereby reaching solid conclusions, (c) they have to figure out the materials or devices required for the conduction of the experiment, (d) propose the appropriate data measuring procedure, and (e) they have to provide criteria for assessing the confirmation (or not) of their hypothesis (Efstathiou et al., 2018; Lefkos et al., 2011; van Riesen et al., 2018b).

Challenges in the DoE

Research indicates that designing an experiment poses significant challenges not only for primary and secondary education students but also for university students (Dasgupta et al., 2014; Deane et al., 2014; Koretsky et al., 2008; Yang and Park, 2017; van Riesen et al., 2018a; van Riesen et al., 2018b). Moreover, undergraduate students, including future teachers, face challenges with DoE even after completing related laboratory courses (Aydoğdu, 2015; Bitzenbauer and Meyn, 2021; García-Carmona et al., 2024).

These challenges are often linked to students holding non-expert-like ideas, commonly referred to as alternative conceptions (AC) or naive/inaccurate conceptions (Brownell et al., 2014; Deane et al., 2014). Research indicates that a significant portion of preservice (future) primary school teachers hold AC, particularly in the field of physics (Métioui, 2023; Stefanidou et al., 2019). For instance, Koc and Yager (2016) identified that 67.4% of preservice elementary teachers held AC concerning fundamental physical science concepts. AC often arise from everyday experiences, prior schooling, and intuitive reasoning, and are not easily corrected by traditional science courses (Koc and Yager, 2016).

Key areas of difficulty include formulating a hypothesis and manipulating variables (Aydoğdu, 2015; Boudreaux et al., 2008; García-Carmona et al., Lawson, 2002; Lefkos, 2024; van Riesen et al., 2018b), or comprehending the problem and interpreting the data (Dasgupta et al., 2014). It has been found that a major reason for their difficulties in creating hypotheses or choosing effective experimental strategies is the limited prior knowledge or insufficient understanding of the subject matter (Dasgupta et al., 2014; Kurup et al., 2019; Schreiber et al., 2012; van Riesen et al., 2018a).

Concerning the hypothesis, the insufficient understanding of the scientific content has been proposed as one of the main reasons for students' failure (Dasgupta et al., 2014; Schreiber et al., 2012).

A hypothesis is an informed prediction about the relationship between the dependent and independent variables, which is a justified answer to the research question (Eastwell, 2014). In a systematic review concerning the challenges in understanding and performing experiments, Kranz et al. (2023) reported that university students face challenges when generating hypotheses. For example, they do not know what a hypothesis is, they work without a hypothesis, they operate abductively, and they only generate hypotheses that seem plausible to them (Kranz et al., 2023).

Students are also facing difficulties associated with the variables of a problem they are required to manipulate, as recently reported in Kranz et al.'s (2023) review. For an experiment to be scientifically valid, the students should follow a Control-of-Variables Strategy (CVS) (Chen and Klahr, 1999), where the value of the independent variable is altered to investigate its potential effect on the dependent variable, while all other involved variables need to be held constant, ensuring that conclusions are derived from unconfounded experiments (Schwichow et al., 2022; van Riesen et al., 2019). However, students often fail to properly apply CVS, often changing too many variables at once, which leads to a "confounding" error (Dasgupta et al., 2016; Schwichow et al., 2022; van Riesen et al., 2019). Hence, the complexity of the experimental task can be a critical factor; tasks related to everyday life problems, particularly those with many variables to be manipulated or resulting in counterintuitive outcomes, require significant cognitive effort (Lefkos, 2024). It has also been found that certain design mistakes related to faulty CVSs are based on students' limited metacognitive knowledge regarding when and why to apply the required scientific rules (Schwichow et al., 2022).

Overall, students often find it challenging to link experimental evidence with relevant theoretical concepts when designing experiments (Psillos, 2023). The inability to transfer knowledge and skills suggests that students may struggle to identify common theoretical features when the experimental context changes, which is a challenge that distinguishes novices from experts (Dasgupta et al., 2014). In addition, when students encounter scenarios where the experimental results contradict their expectations or intuition, their AC tend to emerge and can interfere with the design process (Boudreaux et al., 2008).

Study Aim and Research Questions

In the domain of Science Education, research on various aspects of experimental design ability at the higher education level primarily focuses on students from polytechnic or science faculties, while research in pedagogical departments is limited (Boudreaux et al., 2008; García-Carmona et al., 2024; Kalthoff et al., 2018). Previous studies indicate that students' challenges are connected to insufficient explicit instruction on SPSs and the prevalence of traditional instructional approaches in their university education (Aydoğdu, 2015).

In the present study, we are focused on FPTs. Research indicates that school students begin to develop the ability to design experiments as early as primary school (Osterhaus et al., 2015), making it particularly important to target this audience of FPTs. In addition, providing solid training to future teachers in inquiry-based approaches might result in a wider adaptation of inquiry in school settings (Molohidis and Hatzikraniotis, 2018).

In this paper, we report on research concerning the DoE assessment and the challenges faced by university students/prospective teachers. In addition, we investigate the factors that may contribute to the difficulties encountered. We specifically investigate features contributing to these difficulties by exploring their possible link to: (a) the students' AC, and (b) the complexity of the phenomena under study, defined in terms of the number of variables they have to control. This link has not been explored explicitly for this kind of population.

The research questions of our research are the following:

When future teachers are asked to design experiments based on everyday life problems:

- (RQ1) Do they face any challenges? How is this related to the content and characteristics of the problems they are working on?
- (RQ2) Is their success influenced by the outcome of the experiment? Is this related to their AC?
- (RQ3) Is their success influenced by the complexity of the experiment? Is this related to the number and type of variables they have to manipulate?

METHODOLOGY

The present research employed a quantitative content analysis of student-written experimental designs combined with correlational analysis to examine how cognitive factors (AC) and task complexity (number of variables) relate to future teachers' success in designing experiments (Dasgupta et al., 2014; Yang and Park, 2017).

Sample and Conditions of the Research

The participants in this study were 148 FPTs, all senior students in their sixth semester enrolled in an optional physics course on Science Education. This specific group was selected because mastery of fundamental SPSs, such as identifying variables and interpreting data, is regarded as a core competence necessary for effective elementary science teaching (García-Carmona et al., 2024). In addition, previous studies consistently show that FPTs often begin their training with limited scientific competence (García-Carmona et al., 2024), exhibiting low levels of integrated SPSs (Aydoğdu, 2015; García-Carmona et al., 2024), particularly in complex skills such as hypothesis formulation and controlling variables (Aydoğdu, 2015).

By focusing on FPTs, we aimed to assess their abilities after several years of exposure to science, seeking to identify

persistent difficulties that call for specific interventions in teacher preparation programs (Aydoğdu, 2015). Addressing these deficiencies is essential, as the lack of sufficient preparation may limit a teacher's capacity to create quality learning situations (García-Carmona et al., 2024) and hinder the subsequent adaptation of IBL in future school settings.

Throughout the course, students become familiar with fundamental principles of physics, including heat radiation and water solutions. They are also exposed to methodological approaches, such as designing investigations in the context of IBL, during which individual aspects, such as designing experiments, manipulating variables, and forming hypotheses are studied. During the course, students had to submit several weekly assignments. Some of them were related to IBL and the DoE.

The data herein are collected from a weekly assignment consisting of four (4) diagnostic worksheets submitted by students at the semester's initial stages during an introduction to learning theories and how they are applied in physics education. Therefore, data were collected before discussions about the inquiry-based approach, experimental investigations, and the scientific content, such as the concepts and phenomena mentioned in the worksheets. The purpose of this assignment was to record students' initial views so that they could reflect on them at a later stage.

The data for the present study are related to the assignments submitted by students taking the course in the spring semester, so this is considered a "convenient" sample. The responses were anonymized. Hence, there was no way to identify the participants, their names, their ages, or any other characteristic. Of the 160 assignments collected, 148 responses were considered valid for the analysis described below. Specifically, in worksheet A - 34 responses; in worksheet B - 37 responses; in worksheet C - 37 responses; and in worksheet D - 40 responses. The rest of the responses were considered invalid since they were off-topic or had many missing or one-word statements.

The Structure and Rationale of the DoE Worksheet (DoEW)

Data were collected from the assignment worksheets, in which students filled out their DoE (DoEW), by following some simple open-ended guiding questions. These questions essentially corresponded to the dimensions of the rubric's assessment schema (see the following section) and acted as a scaffold for the students (van Riesen et al., 2019).

This option of utilising structured yet open-ended questions balances, on the one hand, the autonomy of the participants in their responses while, on the other hand, providing the researcher with a tool to collect data and compare the results (Emden and Sumfleth, 2016). Furthermore, it provides the benefit of straightforward and quick integration within the educational process, saving valuable instructional time and requiring no specialized equipment, factors which are considered critical for its acceptance by educators (Ha and Kim, 2020). Last but not least, such a DoEW is versatile and can be applied in any educational format, whether in-person or remote education.

The choice of everyday-life problems involving complex or counterintuitive concepts was utilized to surface students' relevant scientific knowledge and, importantly, their pre-existing non-expert-like beliefs or AC. Moreover, these phenomena are deliberately chosen to increase cognitive demands and challenge students' application of core scientific strategies, such as the CSV.

The structure of DoEW served as a scaffolding tool (van Riesen et al., 2019), breaking down the complex DoE task into manageable, sequential phases, thereby shifting instruction away from traditional "cookbook" activities (Agustian et al., 2022; Aydoğdu, 2015). Students were explicitly required to complete some essential steps of the process, like the following:

formulate a hypothesis by making an informed prediction that links variables; describe the setup, variables (independent, dependent), and procedures, which requires students to operate at higher cognitive levels (analysis, synthesis) (Koretsky et al., 2008); specify criteria for hypothesis confirmation/rejection, guiding students to connect experimental evidence back to the theoretical concept. This structure ensured that the resulting students' designs contained all necessary information for the content analysis following the DoE assessment rubric (see the following section) (Lefkos, 2024).

The Content of the Everyday-life Problems

Following the guidelines of the DoEW, students were given a problem (Table 1), for which they had to express a hypothesis and then describe (design) a valid experiment for testing their hypothesis.

The problems are described in everyday life terms and were chosen appropriately so that, on the one hand, the scientific content was relevant to the students' everyday experience, and on the other hand, the design of the experiments is feasible with reference only to everyday life materials (i.e., no need for specialized instruments or apparatuses). For example, the problem in DoEW-A dealt with the emission of thermal

Table 1: The problems, their content, and their features corresponding to each of the Experiment Design Worksheets (DoEW)

DoEW	Physics content	Feature-1	Feature-2
DoEW-A	Absorption of thermal radiation in correlation to the colour of an object	Intuitive experiment	Many variables
DoEW-B	Water-sugar solution in correlation to water temperature	Intuitive experiment	A few variables
DoEW-C	Emission of thermal radiation in correlation to the color of an object	Counterintuitive experiment	Many variables
DoEW-D	Water–salt solution in correlation to water temperature	Counterintuitive experiment	A few variables

radiation in correlation with the colour of an object, and it was worded as follows: *"A boy gave his father two identical cups for his coffee as a present, except that one was white, and the other one was black. Which one is better for keeping the coffee hot for a longer time? Is it the black, the white, or is it the same for both?"* The other DoEWs had similar wordings and were dealing with the absorption of thermal radiation in relation to the colour of an object, and the solubility of sugar or salt in correlation to the temperature of water.

Next, using the DoEW, students are asked to formulate a hypothesis and then design (describe) a relevant, valid experiment to test it. In addition, they are asked to specify the criterion by which they will decide whether their hypothesis was confirmed or rejected.

The present research is focused on investigating two factors that empirical research has indicated as challenges students encounter when asked to design experiments: formulating a hypothesis and manipulating variables.

For this reason, the problems to which the students were asked to respond in their assignments were deliberately chosen to serve the needs of the research, as they gave us the opportunity to investigate them. At the same time, we attempted to link these factors to the content and specific characteristics of the problems.

In particular, experiments in DoEW-A and B have an outcome, which is related to the existence of alternative views formed based on everyday experience in relevant phenomena. In contrast, experiments in DoEW-C and D have a counterintuitive outcome (Table 1). For example, in the problem of DoEW-C, most participants use their everyday experience of radiation absorption, where black bodies absorb more efficiently (e.g. when someone is wearing black clothes), and assume that the emission effect is inverse; hence they usually answer that the white cup will cool down faster - which is not the case. Similarly, in the experiment of DoEW-D, participants use their experience of the effect of water temperature on the solubility of sugar, which more than doubles in hot water (From 20 to 100°C change 178%), and transfer this experience to table salt, which is not valid.

At the same time, experiments in DoEW-B and D have a small number of variables that students need to manipulate, while experiments in DoEW-A and C have a large number of variables (Table 1). Therefore, this will allow us to test whether these two factors influence the design of the experiments. Specifically, in the experiments of DoEW-B and D, participants have to identify and manipulate one independent variable (amount of sugar/salt), while controlling other variables such as amount of water and water temperature, and measuring the dependent variable (e.g., the amount of sugar/salt dissolved or dissolution rate). On the other hand, in problems A and C, they have to identify and manipulate one independent variable (e.g., type/color of cup) while controlling others (e.g., amount of liquid, ambient temperature, initial temperature), and measuring the dependent variable (e.g., coffee/soda temperature change over a specific time).

The Assessment Tool

Various methodologies have been proposed to evaluate the DoE, such as interviews (Schneider and Bullock, 2011), questionnaires (Efstathiou et al., 2018; Kalthoff et al., 2018; Osterhaus et al., 2015), or rubrics. The application of a rubric for evaluating the DoE provides a distinct advantage compared to alternative methods, serving both as a diagnostic and formative tool to assist participants in their skill development (Etkina et al., 2006), as it provides a descriptive way of analysing the levels of success.

Of the various rubrics that have been suggested by other researchers for university students, none were found that were relatively close to the requirements of this study. For example, in the one proposed by Komives et al. (2007), the participants were from the Faculty of Engineering, and another one was addressed to students from the Faculty of Biology (Dasgupta et al., 2014). Both of them were considered quite complex for our aim.

Consequently, for the present study, we slightly modified and used a rubric that has been proposed to assess the DoE by primary and secondary school students (Lefkos et al., 2011), as it was considered closer to the participants' profiles and capabilities. The modified rubric, adapted to the conditions (age/experience/cognitive level of participants, different educational context) of this particular research, while maintaining its general philosophy, has already been tested in a pilot study with fruitful results (Lefkos, 2024).

In general, the rubric suggests evaluating the DoE along several axes we call "Dimensions," which correspond practically to the different phases of the DoE (Efstathiou et al., 2018; van Riesen et al., 2018b). Each dimension is graded in three (3) predefined levels of success (Appendix A). The lowest success level receives 1 point, and the highest receives 3 points.

To assess the students' level of success in each of the Dimensions, content analysis is used based on the answers given by the participants in the respective worksheet.

The applied rubric comprises 6 dimensions. Table 2 presents the dimensions of the experiment design that are scored based on the rubric used in this research.

Therefore, according to the rubric, the design of an experiment is scored against 6 dimensions and 3 levels of success for each dimension, each of which is assigned 1–3 points.

Table 2: The six dimensions of the design of experiment, assessed by the rubric

Dimension	Description
D1	Formulation of a hypothesis
D2	Criteria for the evaluation of the hypothesis
D3	Recommendation of appropriate materials and apparatuses
D4	Identification and manipulation of variables
D5	Setting of the initial conditions
D6	Description of the experimentation procedure and data collection

Consequently, each dimension can receive a minimum of one (1) point and a maximum of up to three (3) points from each participant (e.g., for each dimension in Essay A, the minimum is 6 × 1 = 6 points, while the maximum is 6 × 3 = 18 points).

As in the present research, we are interested in investigating the possible difficulties encountered by students in designing experiments as a whole and not individually; the focus of the investigation is on the scores of each dimension of experimentation (as formed, of course, after scoring the designs of each participant). For each dimension, the sum of the scores is therefore calculated, and to make comparisons between experiments, the scores are then converted into percentages (%). Thus, dimensions in which participants accomplish low scores mean that they present greater difficulty for them compared to other dimensions in which they accomplish higher scores.

The modified rubric was tested in a preliminary pilot study with a small group of future teachers (Lefkos, 2024). This informed the final refinement of the definitions and scoring criteria, ensuring they were suitable for the participants' cognitive level and experience.

In addition, to ensure methodological rigour and reproducibility, the scoring of the collected data (i.e., the content analysis) was performed by the author and two coders who are specialists in science education. One of the coders is an experienced researcher and academic, while the other is a new researcher/ secondary school teacher.

The group initially agreed on the appropriate responses for the higher level of success in each dimension and then agreed on the responses that would be classified in the other two lower levels of success. This was done separately for each of the four experiments.

High inter-rater reliability is crucial for rubric-based assessment (Dasgupta et al., 2014; Lefkos et al., 2011). All three coders individually assessed 10 DoEW from each experiment and then discussed their assessments to resolve any points of inconsistency. After agreeing on 82.5% of the scoring, the author coded and assessed the remaining designs.

RESULTS

Figures 1 and 2 display the scoring results for each dimension of the DoE (in %) across all experiments. Overall, it is evident that, with a few exceptions, the participants' performance across most dimensions and all tests was notably high (ranging from 60% to 80%).

At the same time, however, we can observe some similarities and differences between the four experiments. In particular, we should note the following:

(a) The success rates of D1 in experiments C and D are analogs to one another (39% and 45% success, respectively), while at the same time, the same dimension in experiments A and B present very high success rates (78% and 84% respectively)

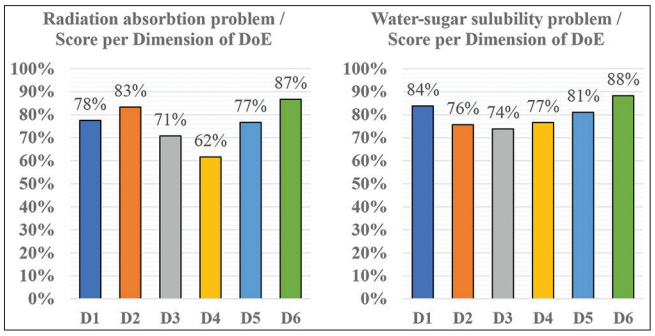

Figure 1: The mean success rates (%) from scoring the six dimensions of design of experiment: Worksheets A and B

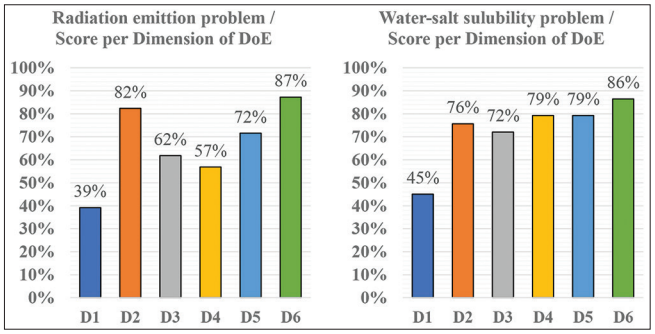

Figure 2: The mean success rates (%) from scoring the six dimensions of design of experiment: Worksheets C and D

(b) In addition, the lower success rates of D4 in experiments A and C are similar (62% and 57%, respectively), while the same dimension in experiments B and D presents very high success rates (77% and 79%, respectively).

In other words, we observe some similarities and some differences in the success rates of students in Dimensions D1 and D4 among the different experiments. As a reminder, D1 is related to the formation of a hypothesis, and D4 is related to the manipulation of the variables.

A statistical analysis follows in the next section to check whether these differences are statistically significant.

Statistical Analysis of the Data

To examine the potential relationships between the dependent variable (Dimensions D1-D6) and the independent variables (the problem characteristics, i.e., their variable complexity and their intuitive outcome), the following steps were followed. For the analysis, SPSS v29 was used.

(i) As a first step, we tested the normality of the distribution of the values in the independent variables. The null hypothesis is H0 = Our data come from a population that follows a normal distribution.

As shown in Table 3, the null hypothesis is rejected in all cases ($p < 0.001$). Therefore, our data do not follow a normal distribution.

(ii) In the next phase, the correlation between the independent variable (dimensions) and the overall characteristics of

the experiment was investigated. In other words, whether students responded to all dimensions the same way in all experiments or not. As the data distribution is not normal, non-parametric independent-sample tests were deployed. In this investigation, the null hypothesis is: H0 = The distribution of Dimension-X is the same in all the experiments.

As shown in Table 4, four of the six dimensions, namely D2, D3, D5, and D6, seem to satisfy the null hypothesis H0. This means that the students respond similarly, without any statistically significant difference ($p > 0.05$) between the different designs of experiment in relation to these Dimensions.

Dimensions D1 and D4, on the other hand, reject the null hypothesis H0. This, therefore, means that the students' responses show a statistically significant difference ($p < 0.001$) between the four designs of the experiment in relation to Dimensions D1 and D4.

This also corresponds to one of our research questions (RQ1) and confirms the research design and our choices regarding the content and characteristics of the problems. Future teachers seem to face challenges when asked to design experiments based on everyday life problems in some aspects of the design. Moreover, these challenges are related to the content and characteristics of the problems they are working on.

(iii) Finally, we further deepened our investigation to search for any possible correlations between the independent variable (dimensions of DoE) and the two dependent variables (the variable complexity and their intuitive/counterintuitive outcome), which comprise the specific characteristics of the problems.

Table 5 presents the findings of Pearson's test, indicating a strong positive correlation between dimension D1 and the intuitive outcome of the problem ($r[146] = 0.655$, $p < 0.001$). A moderate negative correlation was identified between dimension D4 and the number of variables involved in the problem ($r[146] = -0.402$, $p < 0.001$). No statistically significant correlation was observed in any other dimension.

That means that if the outcome of a problem is intuitive, students are more likely to score higher at the "Formulation of a Hypothesis" (D1).

On the other hand, when the problem is more complicated and requires controlling more variables, students are more likely to score lower at the "Identification and manipulation of variables" (D4).

Consequently, regarding our research questions RQ2-RQ3, our data confirm that the success of the future teachers' DoE is influenced by (a) the outcome and (b) by the complexity of the experiment. Specifically, students get a higher score at the "Formulation of a Hypothesis" (D1) when the outcome of the experiment is intuitive, but a lower score when the outcome is counterintuitive.

In addition, students get a lower score in the "Identification and manipulation of variables" (D4) when the problems involve many variables and are thus more complex, but they get a higher score when they are less complex and involve fewer variables.

Table 3: The test for normality of the data distribution

DoE Dimension	Kolmogorov–Smirnov[a]			Shapiro–Wilk		
	Statistic	df	Significant	Statistic	df	Significant
D1	0.297	148	<0.001	0.747	148	<0.001
D2	0.308	148	<0.001	0.756	148	<0.001
D3	0.279	148	<0.001	0.797	148	<0.001
D4	0.265	148	<0.001	0.803	148	<0.001
D5	0.298	148	<0.001	0.768	148	<0.001
D6	0.402	148	<0.001	0.651	148	<0.001

[a]Lilliefors significance correction

DISCUSSION

In this study, the ability of university students – future teachers – to design experiments was investigated based on the data collected from their weekly assignments. Specifically, students' success was evaluated across six (6) dimensions of

Table 4: Investigation of the dependence of students' scores at the six dimensions (D1-D6) on the content of the four different experiments

DoE Dimension	Null hypothesis	Test	Significant[a, b]	Decision
D1	The distribution of D1 is the same across categories of Experiment.	Independent-Samples Kruskal–Wallis Test	<0.001	Reject the null hypothesis.
D2	The distribution of D2 is the same across categories of Experiment.	Independent-Samples Kruskal–Wallis Test	0.260	Retain the null hypothesis.
D3	The distribution of D3 is the same across categories of Experiment.	Independent-Samples Kruskal–Wallis Test	0.129	Retain the null hypothesis.
D4	The distribution of D4 is the same across categories of Experiment.	Independent-Samples Kruskal–Wallis Test	<0.001	Reject the null hypothesis.
D5	The distribution of D5 is the same across categories of Experiment.	Independent-Samples Kruskal–Wallis Test	0.484	Retain the null hypothesis.
D6	The distribution of D6 is the same across categories of Experiment.	Independent-Samples Kruskal–Wallis Test	0.961	Retain the null hypothesis.

[a]The significance level is 0.050, [b]Asymptotic significance is displayed

Table 5: The results of the Pearson correlations between the dimensions of design of the experiment and the characteristics of the problems (variable complexity and intuitive outcome)

Characteristics of the problems	D1	D2	D3	D4	D5	D6
Intuitive output						
Pearson correlation	0.655**	0.017	0.115	0.006	0.066	0.017
Sig. (2-tailed)	<0.001	0.834	0.163	0.940	0.425	0.839
n	148	148	148	148	148	148
Variable complexity						
Pearson Correlation	−0.077	0.161	−0.141	−0.402**	−0.121	−0.013
Sig. (2-tailed)	0.351	0.051	0.088	<0.001	0.143	0.877
n	148	148	148	148	148	148

**Correlation is significant at the 0.01 level (2-tailed)

the DoE: the formulation of a hypothesis about the outcome of the phenomenon, the criteria for the hypothesis evaluation, the identification and manipulation of the variables influencing the phenomenon to arrive at valid conclusions, recommendations for necessary materials and apparatus, the setting of initial conditions, and the description of a valid experimentation procedure and data collection. This investigation showed that participants had fairly high success rates in most of the above dimensions, while they had difficulties in others, such as hypothesis formulation and variable manipulation.

To accomplish the above investigation, we used an evaluation rubric comprising six (6) axes, the above-mentioned as dimensions, and three (3) levels of success in each. This rubric was a modified version of one used previously for the same purpose with high school students (Lefkos et al., 2011), but was also piloted earlier with a few future teachers (Lefkos, 2024).

The rubric was utilized with an appropriately structured DoEW, thanks to which data were collected from students' assignments. Then, using a content analysis methodology, the designs were evaluated based on the different levels of success of each dimension. The DoEW contributed to the collection of the responses but also to their assessment in a way that is as reliable as possible (Emden and Sumfleth, 2016), as the structural elements of the DoEW were juxtaposed with the six dimensions of the DoE.

Each of the four DoEW in the assignments contained a different everyday-life problem that the future teachers were asked to transform into a valid experiment. Certain features of these problems facilitated the identification of the participants' potential difficulties.

From previous Science Education research, either with students or future teachers, it is known that challenges in DoE occur, particularly in formulating hypotheses or managing variables (Boudreaux et al., 2008; Lawson, 2002; van Riesen et al., 2018b), which also aligned with the results of our study. The challenges have been attributed to an insufficient understanding of the scientific content (Dasgupta et al., 2014; Schreiber et al., 2012) or to the various types of variables participants have to handle (Arnold et al., 2014; Dasgupta et al., 2014). However, the effect on the DoE of (a) the existence of AC in participants and (b) the multitude of variables they are called to manipulate, which we found in the present study, has not been particularly explored, at least for the case of FPTs, as in our case.

In our view, the apparent correlation of challenges in the DoE with the above-mentioned characteristics of the problems that may be included in the inquiry-based activities is particularly important for the university educators who design these activities. For example, when an educator is aware of the above findings and wishes to design a diagnostic activity, he or she could include problems of increased difficulty (e.g., problems with many variables and non-intuitive outcomes) or of lesser difficulty (e.g., problems with few variables and intuitive outcome), depending on the ability level of the participants and the objectives of the activity. Of course, the possible identification of these difficulties also points to the need to reinforce participants, either at the level of scientific content and/or at the level of strategies for variable manipulation. Accordingly, in activities aiming to develop the participants' DoE ability, one might start with problems of low difficulty and gradually increase their difficulty while training them both in terms of the scientific content and the strategies for manipulating variables.

Counterintuitive Outcomes and the Formulation of Hypothesis (RQ2)

The finding that students score lower in "Formulation of a Hypothesis" (D1) when the experiment outcome is counter-intuitive directly links their success to the influence of their AC or prior knowledge structures (Brownell et al., 2014; Deane et al., 2014). When the expected result conflicts with intuition, AC seem to interfere with the design process. This outcome aligns with research confirming that inadequate prior knowledge or conceptual understanding is a key factor in failure to formulate hypotheses and shows that DoE success is not only a procedural skill but is fundamentally connected with the domain-specific content knowledge matter (Dasgupta et al., 2014; Kurup et al., 2019; Schreiber et al., 2012; van Riesen et al., 2018a).

Complexity of the Problems and Manipulation of Variables (RQ3)

The finding that students score lower in "Identification and Manipulation of Variables" (D4) when problems are more complex (requiring the manipulation of numerous variables) directly confirms their difficulty with applying the CVS. This challenge of manipulating too many variables simultaneously or failing to account for all factors affecting a phenomenon is a well-documented problem observed across various educational levels (Dasgupta et al., 2014; Schwichow et al., 2022).

DoE as a Distinct Phase of IBL

One of this paper's theoretical assumptions is that designing an experiment can be considered a distinct, stand-alone phase

of experimentation (Kipnis and Hofstein, 2008) and practising inquiry in educational settings. In this line of rationale, the importance of having an assessment rubric, which, in combination with the DoEW, can be easily and quickly implemented, is very important for educators, as it offers the possibility to implement experiment design activities with their students, without the need for special equipment and in a shorter time (Ha and Kim, 2020).

At the same time, we consider that it also offers the advantage of being applicable to different educational scenarios. For example, an educator can implement this kind of design activity either linked to the actual execution of the experiment or just as a stand-alone activity. While in the first case, students should be provided with all necessary materials and equipment to execute the experiment, the distinct relationship between the design phase, as adopted in this paper, and the experimental phase offers the additional advantage of implementing the relevant activities, regardless of the experimental environment, whether it is a physical or a virtual laboratory, in face-to-face or distance learning conditions.

On the other hand, the latter case of a stand-alone activity offers even greater application flexibility in various educational settings as a paper-and-pencil activity.

The Proposed Assessment Rubric

Teachers, in general, in their challenging role of planning activities and assessing students, need easy-to-use and flexible tools. An assessment rubric such as the one used in this paper, combined with the DoEW, provides the means for easy implementation of relevant design activities. As these activities can be implemented without the need for laboratory equipment, independent of the execution of an experiment, the teacher is given the possibility to use them flexibly, either for the diagnosis of the experimental design ability of the students or as a tool for formative assessment and contribution to the development of this ability. The rubric's diagnostic capabilities can help teachers to identify specific student misconceptions early, enabling timely and targeted instructional adjustments, thus maximising teaching efficiency.

Future teachers will eventually be asked to take on this role later in their professional careers. It is therefore important that their training at the university provides them with an appropriate pedagogical background and introduces them to easy-to-use tools aimed at acquiring a positive attitude towards inquiry-based approaches. These can develop their future students' problem-solving skills, which seems to be important both at the personal and professional levels.

Implications for Science Education Research and Teacher Education

The findings of this study provide critical implications for both the curriculum design and research agenda within Science Education, particularly concerning future teachers. Our research can contribute to changing the perspective of (future) educators in favour of adopting inquiry-based practices and implementing corresponding design activities. It has been found that teachers are reluctant to implement inquiry-based approaches (Melville et al., 2013), claiming either the lack of laboratory equipment or the lack of teaching time (Ha and Kim, 2020). However, another reason is their lack of familiarity, and the training of future teachers at the university could contribute toward this (Molohidis and Hatzikraniotis, 2018). Moreover, as our study found, FPTs struggle with essential inquiry practices due to weaknesses in content knowledge and metacognitive skills, thus making them less willing to implement IBL effectively in their future classrooms (Aydoğdu, 2015; Garbett, 2003; García-Carmona et al., 2024). Teaching science successfully requires both content knowledge and pedagogical content knowledge (García-Carmona et al., 2024).

The solution is not just about doing "more labs," but about implementing explicit instruction that fosters higher-order experimental competences (Boudreaux et al., 2008; Schwichow et al., 2022). Teacher education needs to move away from prescriptive, traditional "cookbook" experiments toward guided, reflective inquiry (Agustian et al., 2022; Aydoğdu, 2015; van Riesen et al., 2018a).

Activities such as designing experiments can promote aspects of inquiry that are sometimes considered more important than the actual conducting of experiments (Garratt and Tomlinson, 2001). Such activities can both promote the application of inquiry-based approaches and be easily implemented in a variety of educational settings. Therefore, we believe that it is very useful for future teachers to experience them, to practice on them, and to discuss their educational implications.

Study Limitations

Our study has several limitations that should be considered when interpreting the findings. First, the use of a convenience sample from a single institution limits the generalizability of the results to broader populations of future teachers, contrasting with larger, multi-institutional, or systematic investigations (e.g., Kranz et al., 2023). Second, due to the fact that data were collected before any formal instruction on experimental design, the results provide an authentic baseline of pre-instructional abilities but cannot make inferences about how explicit pedagogical interventions might influence FPTs' learning gains. Hence, future studies could adopt quasi-experimental designs, as implemented in related research (e.g., Dasgupta et al., 2014; van Riesen et al., 2018b), to compare pre- and post-instructional performance. Finally, although inter-rater reliability was initially established for more than 25% of the entries, most of the data were coded by a single author, which may introduce potential bias. Future research could employ multiple coders of the entire dataset to enhance methodological rigour and strengthen the validity of findings.

CONCLUSIONS

In this paper, we studied the challenges faced by FPTs when designing experiments for everyday phenomena with complex and counterintuitive outcomes. We identified two key factors

that contribute to suboptimal performance in DoE: (a) the conflict between AC and expected outcomes, which mainly hinders hypothesis formulation, and (b) the cognitive difficulty caused by the complexity of tasks with many variables, thus highlighting the challenges in applying the CVS.

These findings enrich the field of science education, confirming the complex relationship between cognitive conceptual knowledge (CK) and experimental procedural competence, especially in the important population of FPTs (Agustian et al., 2022; van Riesen et al., 2018a). The identified gaps demonstrate that current pedagogical approaches are insufficient for cultivating higher-level experimental skills and metacognitive awareness, which are necessary for effective science teaching (Agustian et al., 2022; Aydoğdu, 2015; Schwichow et al., 2022).

To address these gaps and align Teacher Education and Training with the goals of research-oriented curricula, such as K-12 standards (Schwichow et al., 2022), a shift of focus in two directions might be implemented, as proposed below.

The systematic integration of explicit instruction in FPTs training, focusing on experimental design and metacognitive strategies, into IBL strategies, with a gradual increase in complexity (Agustian et al., 2022; van Riesen et al., 2019).

Future research should investigate the effectiveness of instructional modules that incorporate model-based reasoning frameworks (e.g., as conceptualized in Boudreaux et al., 2008; Dasgupta et al., 2014; Lawson, 2002) to support hypothesis formation, as well as explicit guidance for applying the CVS (e.g., following instructional models from van Riesen et al., 2018b or meta-analyses from Schwichow et al., 2022) when FPTs encounter counterintuitive or cognitively demanding problems.

This work could benefit from design-based research methodologies (e.g., as applied in Agustian et al., 2022) to iteratively improve such instructional interventions, followed by quasi-experimental studies to rigorously assess their impact on FPTs' experimental design abilities.

The use of diagnostic rubrics or other methodologies (e.g., think-aloud protocols, observations) to identify error patterns and analyse procedural and metacognitive struggles of FpTs, with a particular focus on the transfer of scientific skills to different conceptual domains (Brownell et al., 2014; Psillos, 2023; Schwichow et al., 2022; van Riesen et al., 2019). Controlled experimental studies, similar to previous intervention research (e.g., van Riesen et al., 2018b), could assess the effectiveness of these diagnostic tools in identifying students' difficulties, for example, investigate metacognitive struggles that FPTs experience when manipulating variables in complex experimental designs or examine which instructional scaffolds effectively address these difficulties to improve the application of the CVS and hypothesis formulation. Such studies would provide actionable insights for designing targeted interventions to overcome persistent barriers in FPTs' experimental design proficiency.

This way, the present study provides a framework for the development of targeted interventions that will enhance the professional competence of future teachers, facilitating the meaningful adoption of IBL in educational settings.

ETHICS STATEMENT

When the data were collected, our study met the ethical requirements of the University of Macedonia (Committee for Research Ethics). According to the committee, this research was based on data from secondary sources; hence, the requirement for consent was waived.

REFERENCES

Abd-El-Khalick, F., Boujaoude, S., Duschl, R., Lederman, N.G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H.L. (2004). Inquiry in science education: International perspectives. *Science Education*, 88(3), 397-419.

Agustian, H.Y., Finne, L.T., Jørgensen, J.T., Pedersen, M.I., Christiansen, F.V., Gammelgaard, B., & Nielsen, J.A. (2022). Learning outcomes of university chemistry teaching in laboratories: A systematic review of empirical literature. *Review of Education*, 10, e3360.

Arnold, J.C., Kremer, K., & Mayer, J. (2014). Understanding students' experiments-what kind of support do they need in inquiry tasks? *International Journal of Science Education*, 36(16), 2719-2749.

Aydoğdu, B. (2015). Examining preservice science teachers' skills of formulating hypotheses and identifying variables. *Asia-Pacific Forum on Science Learning and Teaching*, 16(1), 1-38.

Bitzenbauer, P., & Meyn, J.P. (2021). Fostering experimental competences of prospective physics teachers. *Physics Education*, 56(4), 045020.

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

Brownell, S.E., Wenderoth, M.P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., Walcher-Chevillet, C.L., & Crowe, A.J. (2014). How students think about experimental design: Novel conceptions revealed by in-class activities. *BioScience*, 64(2), 125-137.

Cai, B., Mainhood, L., Groome, R., Laverty, C., & Mclean, A. (2021). Student behavior in undergraduate physics laboratories: Designing experiments. *Physical Review Physics Education Research*, 17(2), 020109.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098-1120.

Dasgupta, A.P., Anderson, T.R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE-Life Sciences Education*, 13(2), 265-284.

Dasgupta, A.P., Anderson, T.R., & Pelaez, N.J. (2016). Development of the neuron assessment for measuring biology students' use of experimental design concepts and representations. *CBE Life Sciences Education*, 15(2), ar10.

Deane, T., Nomme, K., Jeffery, E., Pollock, C., & Birol, G. (2014). Development of the biological experimental design concept inventory (BEDCI). *CBE-Life Sciences Education*, 13(3), 540-551.

Eastwell, P. (2014). Understanding hypotheses, predictions, laws, and theories. *Science Education Review*, 13(1), 16-21.

Efstathiou, C., Hovardas, T., Xenofontos, N.A., Zacharia, Z.C., DeJong, T., Anjewierden, A., & Van Riesen, S.A.N. (2018). Providing guidance in virtual lab experimentation: The case of an experiment design tool. *Educational Technology Research and Development*, 66(3), 767-791.

Emden, M., & Sumfleth, E. (2016). Assessing students' experimentation processes in guided inquiry. *International Journal of Science and Mathematics Education*, 14(1), 29-54.

Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D.T., Gentile, M., Murthy, S., Rosengrant, D., & Warren, A. (2006). Scientific abilities and

their assessment. *Physical Review Special Topics - Physics Education Research*, 2(2), 020103.

Garbett, D. (2003). Science education in early childhood teacher education: Putting forward a case to enhance student teachers' confidence and competence. *Research in Science Education*, 33(4), 467-481.

Garratt, J., Tomlinson, J. (2001). Experimental design–can it be taught or learned. *University Chemistry Education*, 5, 74-79.

García-Carmona, A., Muñoz-Franco, G., Criado, A.M., & Cruz-Guzmán, M. (2024). Validation of an instrument for assessing basic science process skills in initial elementary teacher education. *International Journal of Science Education*, 46(4), 362-381.

Girault, I., D'Ham, C., Ney, M., Sanchez, E., & Wajeman, C. (2012). Characterizing the experimental procedure in science laboratories: A preliminary step towards students experimental design. *International Journal of Science Education*, 34(6), 825-854.

Ha, S., & Kim, M. (2020). Challenges of designing and carrying out laboratory experiments about Newton's second law: The case of Korean gifted students. *Science and Education*, 29(5), 1389-1416.

Harlen, W. (2013). Bell, D., Dolin, J., Léna, P., Peers, S., Person, X., Rowell, P., & Saltiel, E. (Eds.). *Assessment and Inquiry-Based Science Education: Issues in Policy and Practice*. Italy: Global Network of Science Academies (IAP).

Hatzikraniotis, E., Kallery, M., Molohidis A., & Psillos, D. (2010). Secondary students' design of experiments after engagement in an innovative inquiry-oriented module on heat transfer. *Physics Education*, 45(4), 335-343.

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.

Karagianni, H., & Psillos, D. (2022). Investigating the effectiveness of explicit and implicit inquiry-oriented instruction on primary students' views about the non-linear nature of inquiry. *International Journal of Science Education*, 44(4), 604-626.

Karelina, A., & Etkina, E. (2007). Acting like a physicist: Student approach study to experimental design. *Physical Review Special Topics Physics Education Research*, 3(2), 020106.

Kipnis, M., & Hofstein, A. (2008). The inquiry laboratory as a source for development of metacognitive skills. *International Journal of Science and Mathematics Education*, 6(3), 601-627.

Koc, I., & Yager, R. (2016). Preservice elementary teachers' alternative conceptions of science. *Cypriot Journal of Educational Sciences*, 11, 144-159.

Komives, C., Mourtos, N.J., McMullin, K.M., & Anagnos, T. (2007). Evaluating Student Mastery of Design of Experiment. In: *37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports*, pp. T3G-7-T3G-12.

Komives, C.F. (2015). Inquiry-based laboratory for teaching students design-of-experiments. *Journal of Engineering Education Transformations*, 28(2), 1-5.

Koretsky, M.D., Amatore, D., Barnes, C., & Kimura, S. (2008). Enhancement of student learning in experimental design using a virtual laboratory. *IEEE Transactions on Education*, 51(1), 76-85.

Kranz, J., Baur, A., & Möller, A. (2023). Learners' challenges in understanding and performing experiments: A systematic review of the literature. *Studies in Science Education*, 59(2), 321-367.

Kurup, P., Li, X., Powell, G., & Brown, M. (2019). Building future primary teachers' capacity in STEM: Based on a platform of beliefs, understandings and intentions. *International Journal of STEM Education*, 6, 1-14.

Lawson, A.E. (2002). Sound and faulty arguments generated by preservice biology teachers when testing hypotheses involving unobservable entities. *Journal of Research in Science Teaching*, 39(3), 237-252.

Lefkos, I. (2024). An assessment rubric for future teachers' ability to design experiments. In: Fazio, C., & Logman, P., (Eds.), *Physics Education Today: Challenges in Physics Education.* Berlin: Springer, pp. 105-117.

Lefkos, I., Psillos, D., & Hatzikraniotis, E. (2011). Designing experiments on thermal interactions by secondary-school students in a simulated laboratory environment. *Research in Science and Technological Education*, 29(2), 189-204.

Melville, W., Bartley, A., & Fazio, X. (2013). Scaffolding the inquiry continuum and the constitution of identity. *International Journal of Science and Mathematics Education*, 11(5), 1255-1273.

Métioui, A. (2023). Primary school preservice teachers' alternative conceptions about light interaction with matter (reflection, refraction, and absorption) and shadow size changes on earth and sun. *Education Sciences*, 13(5), 462.

Millar, R. (2004). The role of practical work in the teaching and learning of science. In: *High School Science Laboratories: Role and Vision*. National Academy of Sciences, p. 25.

Molohidis, A., & Hatzikraniotis, E. (2018). Introducing preservice science teachers in the development of inquiry-based activities. In: *The Role of Laboratory Work in Improving Physics Teaching and Learning*. New York: Springer International Publishing.

Osterhaus, C., Koerber, S., & Sodian, B. (2015). Children's understanding of experimental contrast and experimental control: An inventory for primary school. *Frontline Learning Research*, 3(4), 56-94.

Psillos, D. (2023). The role and impact of virtual laboratories in physics teaching and learning: A synthesis of literature. In: Taşar, M.F., & Heron, P.R.L., (Eds.), *The International Handbook of Physics Education Research: Teaching Physics*. Ch. 2. Melville: AIP Publishing.

Schneider, W., & Bullock, M., (Eds.). (2011). Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study. In: *Human Development from Early Childhood to Early Adulthood: Findings from a 20 Year Longitudinal Study*. England: Psychology Press.

Schreiber, N., Theyssen, H., & Schecker, H. (2012). Experimental competencies in science: A comparison of assessment tools. In: *E-Book Proceedings of the ESERA 2011 Conference: Science Learning and Citizenship. Part 10 Evaluation and Assessment of Student Learning*, pp. 66-72.

Schwichow, M., Brandenburger, M., & Wilbers, J. (2022). Analysis of experimental design errors in elementary school: How do students identify, interpret, and justify controlled and confounded experiments? *International Journal of Science Education*, 44(1), 91-114.

Shanks, R.A., Besterman, H.E., Ziccardi, J.G., & Wozniak, N.M. (2017). Measuring and advancing experimental design ability in an introductory course without altering existing lab curriculum. *Journal of Microbiology and Biology Education*, 18(1), 1-8.

Sokołowska, D. (2018). Effectiveness of learning through guided inquiry. In: *The Role of Laboratory Work in Improving Physics Teaching and Learning*. New York: Springer International Publishing, pp. 243-255.

Stefanidou, C., Tsalapati, K., Ferentinou, A., & Skordoulis, C. (2019). Conceptual difficulties pre-service primary teachers have with static electricity. *Journal of Baltic Science Education*, 18(2), 300-313.

Van Riesen, S.A.N., Gijlers, H., Anjewierden, A., & De Jong, T. (2018a). The influence of prior knowledge on experiment design guidance in a science inquiry context. *International Journal of Science Education*, 40(11), 1327-1344.

Van Riesen, S.A.N., Gijlers, H., Anjewierden, A., & De Jong, T. (2018b). Supporting learners' experiment design. *Educational Technology Research and Development*, 66(2), 475-491.

Van Riesen, S.A.N., Gijlers, H., Anjewierden, A.A., & De Jong, T. (2019). The influence of prior knowledge on the effectiveness of guided experiment design. *Interactive Learning Environments*, 30, 17-33.

Yang, H.G., & Park, J. (2017). Identifying and applying factors considered important in students' experimental design in scientific open inquiry. *Journal of Baltic Science Education*, 16(6), 932-945.

APPENDIX

Appendix A: Dimensions of experiment design assessment rubric

D1 - Statement of Case/Justification

Level Y1	The student expresses assumptions based on intuitive criteria or alternative views
Level Y2	The student expresses hypotheses based on scientifically accepted criteria
Level Y3	The student expresses hypotheses using scientific terminology

D2 - Verification/conclusion

Level E1	The student can specify the verification criterion of his/her hypothesis incompletely/not at all
Level E2	The student defines the criterion based on his/her personal expectations
Level E3	The student can clearly identify the criterion for the verification of his/her hypothesis

D3 - Materials, instruments and devices

Level O1	The student mentions a few of the necessary (or irrelevant to the experiment) materials and apparatuses
Level O2	The student mentions most of the necessary materials and apparatuses
Level O3	The student mentions all the necessary materials and apparatuses

D4 - Identification of variables

Level M1	The student does not identify the dependent and independent variables
Level M2	Student incompletely identifies the dependent and independent variables
Level M3	The student fully identifies the dependent and independent variables

D5 - Setting initial conditions/device settings

Level A1	The student does not specify the initial conditions
Level A2	Student incompletely specifies the initial conditions
Level A3	The student fully specifies the initial conditions

D6 - Experimental procedure/measurement/recording

Level P1	The student cannot describe the process of experimentation and data collection
Level P2	The student vaguely describes the process of experimentation and data collection
Level P3	The student clearly describes the process of experimentation and data collection