Journal of Turkish Science Education, 2024, *21*(3), 484-511.
DOI no: 10.36681/tused.2024.026

Progress in developing experimental design skills among junior high school learners

Luca Szalay[1], Zoltán Tóth[2], Réka Borbás[3], István Füzesi[4]

[1]*Eötvös Loránd University, MTA-ELTE Research Group on Inquiry-Based Chemistry Education, Research Programme for Public Education Development of the Hungarian Academy of Sciences, Hungary, Corresponding author, luca.szalay@ttk.elte.hu, ORCID ID: 0000-0003-0176-0645*
[2]*University of Debrecen, Hungary, ORCID ID: 0009-0000-4806-6840*
[3]*Szent István Secondary School, Hungary, ORCID ID: 0000-0002-9671-087X*
[4]*Eötvös Loránd University, Bolyai János Practicing School, Hungary, ORCID ID: 0000-0003-4826-1819*

ABSTRACT
This paper reports the findings of the second year of a four-year empirical research project. Its aim is to modify 'step-by-step' instructions for practical activities in a way that may enable the development of experimental design skills among junior high school learners. Each school year pupils spend six lessons doing practical activities using worksheets we provide. At the beginning of the research, the Grade 7 (12–13-year-old) pupils were divided into three groups. Group 1 (control group) followed step-by-step instructions. Group 2 followed the same instructions as Group 1, but after the experiment, they answered a series of questions on their worksheets concerned with the design of the experiment. Group 3 was required to design the experiments, guided by a similar set of questions.The impact of the intervention on pupils' experimental design skills (EDS) and disciplinary content knowledge (DCK) was measured using structured tests at the beginning of the project and at the end of both school years. Seven hundred fifty-six (756) Grade 8 pupils completed the test at the end of the second school year (April-May 2023). Over the first two years, the intervention resulted in a medium effect size positive change in the EDS of Group 3 compared to the control group (Group 1), (Cohen's d: 0.23). By the end of the second year of the project, there was only a small difference in the change in DCK between the experimental groups and the control group (Cohen's d value for Group 2: 0.10 and for Group 3: 0.12).

RESEARCH ARTICLE

ARTICLE INFORMATION
Received: 16.08.2023
Accepted: 17.05.2024

KEYWORDS:
Experimental design, inquiry-based education, guided inquiry, chemistry education.

To cite this article: Szalay, L., Tóth, Z., Borbás, R. & Füzesi, I. (2024). Progress in developing experimental design skills among junior high school learners. *Journal of Turkish Science Education, 21*(3), 484-511. DOI no: 10.36681/tused.2024.026

Introduction

Disciplinary Content Knowledge and Inquiry-based Activities

The Rocard Report (2007) recommended inquiry-based methods to increase interest in school science, as policymakers across Europe were concerned that a decline in young people's interest in certain science studies was leading to a shortage of scientists and that not all young people were developing at school the key analytical skills that would prepare them for the future. Inquiry in

science is the intentional process of diagnosing situations, formulating problems, critiquing experiments and distinguishing alternatives, planning investigations, researching conjectures, searching for information, constructing models, debating with peers using evidence, and presenting coherent arguments (Linn, Davis, & Bell, 2004). Banchi and Bell (2008) outline four levels of inquiry. The highest level is the open inquiry, where learners are asked to formulate the research question, design and follow through with a developed procedure, and communicate their findings and results. Open inquiry is certainly authentic, but often considered to be too challenging even at undergraduate level (Farley et al., 2021). One level down is the guided inquiry (Schoffstall and Gaddis, 2007), where the question to be investigated is given by the teacher and learners have to design a procedure to find the answer given guidelines. Guided inquiry is more realistic at school level and it is an improvement on the even lower level of structured inquiry, where the initial question and an outline of the procedure are given to the learners, who are only required to formulate explanations for their finding. The lowest level, which Xu and Talanquer (2013) do not even call an inquiry, is the confirmation inquiry, when the teacher develops the question to be investigated and a procedure that guides the learners through an activity where the results are already known. Banchi and Bell (2008) suggest that teachers should start teaching inquiry at lower levels and work their way up to open inquiry in order to effectively develop students' inquiry skills. The development of skills, habits and attitudes for scientific inquiry is widely seen as an important goal of science education (e.g. Koomson et al., 2024).

However, according to the PISA 2015 results (OECD, 2016), enquiry-based science activities (that are also referred to in the literature as inquiry-based, as the latter term is used in both North American English and British English, see Oxford Learner's Dictionaries, 2024) are associated with lower test scores in science among students who work in the context of disorderly classrooms (Mostafa et. al., 2018). This is unfortunate, because the same analysis also showed that introducing pupils to these activities seems to be the most promising approach to developing a positive attitude towards science (see also e.g. Wheatley, 2018). However, Lavonen et al. (2021) warned that although results show that these science practices have a positive impact on situational interest, several research projects on the topic (European Union, 2016) had not supported the development of students' interest in physics and science careers at upper secondary level. These results suggest that not all inquiry-based teaching methods produce the expected positive results in all circumstances.

Mostafa et al. (2018) also report that greater exposure to teacher-led science education is positively associated with science achievement in almost all countries, even after accounting for socio-demographic characteristics of learners and observed and unobserved school characteristics (OECD, 2016). Therefore, they recommend that teachers with strong classroom management skills and professional knowledge should guide learning in science by explicitly teaching basic concepts. They can then encourage pupils to engage in enquiry or inquiry-based activities to consolidate their knowledge. There are known methods for learners to purposefully integrate scientific knowledge (disciplinary core ideas) and scientific activities (science and engineering practices) to understand aspects of their learning (National Research Council, 2012; NGSS Lead States, 2013). These include e.g. asking questions, developing and using models, designing and carrying out experiments, analysing and interpreting data. This should be productive for making sense of phenomena, while learners adopt epistemologies for science (Russ, 2014), which can turn out to be useful in post-school life. These are skills that all STEM undergraduates should develop (Reynders et al., 2019). Therefore, teaching materials should provide opportunities for learners to surface ideas and build meaningfully on their reflections about the phenomena they experience (Schafer et al., 2023). Exemplars were based on different models, such as the Science Writing Heuristic (Burke et al., 2006), the Process Oriented Guided Inquiry Learning (Moog et al., 2008) and the Argument Driven Inquiry (Walker et al., 2013). These emphasise asking scientific questions, designing appropriate procedures to test those questions, supporting conclusions with experimental evidence, and communicating ideas clearly (Reynders et al., 2019).

In terms of laboratory exercises, it has been known for a long time that step-by-step (or "cookbook-style") instructions that tell students exactly what to do while carrying out an experiment

(which fall under the category of structured inquiry) have limited effects on learning (Kirschner, 1992). This approach usually targets the cognitive, psychomotor and affective domains to ensure meaningful learning. However, the social and epistemic domains are often ignored or dismissed altogether, even in undergraduate laboratories (Hendra, 2022). Therefore, guided inquiry may be an option that represents an improvement over a structured inquiry, but is still less challenging than an open inquiry.

Literature Review

Experimental Design in Pre-university Chemistry Education

There are many different ways to implement a guided inquiry (e.g., Moog et al., 2008), but controlling for variables while designing an experiment is always essential (Arnold et al., 2018; Arnold et al., 2021; Cannady et al., 2019). At school level, where pupils' prior knowledge of chemistry and laboratory competences is extremely limited, it is important to use pedagogical procedures that allow for different levels of support to provide pupils with appropriate scaffolding. Appropriate teaching models can be summarised as not recipe-like instructions, where tasks are more learner-centred, cooperative learning is supported and the preparatory phase plays an important role (van Brederode et al., 2020). The objectives move away from the acquisition of concrete subject content and the use of laboratory equipment towards higher order cognitive skills such as designing experiments, scientific reasoning, linking science and social contexts, and developing critical thinking.

However, Akuma and Callaghan (2019), in their systematic literature review, characterised a number of intrinsic challenges related to the design and implementation of inquiry-based practical work (e.g., negative views about science and practical work, difficulties in designing such activities, persuading learners to reflect on their experiences and outcomes, and concerns about the assessment of practical inquiry). Learners often have to cope with too many instructions in the laboratory (Agustian and Seery, 2017; Johnstone, 1997), and cognitive overload can easily occur. Johnstone (2006) suggested that this could be reduced by pre-laboratory preparation plays. A meta-analysis of studies on guided inquiry instructions suggested that more specific guidance results in higher quality learning products (Lazonder, Harmsen, 2016). Therefore, van Brederode et al. (2020) used two different ways to treat their pre-university students (17–18 year-olds) to find out what level of support works better. In the "critical thinking" pre-laboratory group, students started to develop an experimental plan using the information provided and the criteria for a good experimental design. Hints for designing the experiment were given to the students in the other "paved road" group as information for answering the pre-laboratory questions, while they were also given compact laboratory instructions for carrying out the experiment. The results showed that students of the "critical thinking" group were motivated to think more deeply about the meaning of their measurements than the "paved road" group.

Hennah et al. (2022) found that placing greater emphasis on dialogic processes as a tool prior to completing a practical-based activity helped learners to score significantly higher on the GCSE chemistry examination practical-themed questions than students who prepared for the practical task by watching videos during the lesson. It therefore seems useful for the group of students to discuss the procedure before carrying out a laboratory experiment.

Potier (2023) introduced guided inquiry to enable 15–16 year-old learners with minimal background knowledge of a topic to design and carry out an experiment. He found that guided inquiry can be an effective tool to give students control over their own learning and support their engagement and progress in high school science.

Tseng et al. (2022) argued that since experimental design is a systematic thinking process that involves configuring the relationship between control, and independent and dependent variables (Pedaste et al., 2015), students can learn this by reflective reading of scientific articles rather than performing practical hands-on laboratory activities. In their research project, the "comparison group" (control) learned how to read and understand scientific articles without a direct focus on their inquiry practice. Both "experimental groups" read and discussed scientific articles and designed their own

experiments. Evaluative reflection on their peers' experimental designs was emphasized for one of the experimental groups, while the recognition of variables in designing experimental procedures was emphasized for the other group. The results showed that students' scientific inquiry performance in formulating research questions and designing experimental procedures can be effectively improved by reading and reflecting on experimental design.

Addressing the Problem of Motivation

A systems thinking approach can help learners link their knowledge of chemistry with other disciplines and the skills needed to tackle complex global problems (Szozda et al., 2022). These activities can illustrate e.g. the dynamic relationships between processes, the variables that control them, the emergent behaviour of the system, and how that behaviour changes over time (Orgill, 2019). They provide opportunities for learners to use their knowledge of chemistry to explain a more complex and unfamiliar phenomenon. This can also create an experience that has the potential to engage learners (Allred et al., 2022). This can be very useful, especially in a longitudinal study, when it is crucial to find ways to stimulate and sustain interest, as pupils' performance cannot be rewarded with marks. Several studies have shown that for many high school pupils, motivation to learn chemistry is first and foremost obtaining good marks (see among the latest e.g. Zhang, Zhou, 2023).

Context-based learning in general, and addressing socio-scientific issues in particular, can lead to a better understanding of chemistry and help learners to relate chemistry to their everyday lives (Chen, Xiao, 2020; del Mar López-Fernández et al., 2022). It can also develop critical thinking skills, which are essential for students to become competent citizens who can make informed decisions in different situations. Jiménez-Aleixandre & Erduran (2007) point out that critical thinking involves elements of argumentation, such as the search for and use of evidence. The systems thinking approach is well suited for this purpose, since chemical reactions and processes are an integral part of dynamic and interconnected systems. This way learners can realise that because sustainability has a molecular basis, chemistry plays a central role in addressing the challenges facing the Earth and societal systems (MacDonald et al., 2022). However, to ensure the appropriate development of critical thinking skills, it is important to create a spiral scaffolding while applying a systems thinking approach (Mahaffy et al., 2018). Therefore, at the age of 13–14, phenomena can be explained mostly in qualitative terms and some elements of systems thinking can be introduced. Examples include identifying the components of a system and their connections, flows and cycles, causality and feedback loops (MacDonald et al., 2022). It is best to start simple and then gradually increase the complexity (del Mar López-Fernández et al., 2022).

Social media is also a goldmine to provide topics for motivating context-based learning and systems thinking, as it often contains (e.g. for marketing purposes) science-related information that is based on non-scientifically-proven sources or outright fabrications. Research by Belova & Krause (2023) has shown that it is worthwhile preparing school learners against science-based manipulation strategies. As well as stimulating interest, it can show pupils that seeking to understand how science works can be an activity that protects them from being deceived or misled.

Previous Results

Four studies (Szalay et al., 2020; Szalay et al., 2021; Szalay et al., 2023; Szalay, Tóth, 2016) have provided preliminary results for the research described in this paper. A common feature is that the experimental group or groups learned how to design the experiments, while the control group simply followed step-by-step recipes i.e. structured inquiry. The experimental design tasks of (at least one) of the experimental groups can be categorised as guided inquiry, as the research questions were always given, but the pupils had to design the method, the way to find the answer. The earliest brief empirical research (Szalay, Tóth, 2016), in which pupils in the experimental group aged 14–15 had to design experiments without any help, showed positive results, as their experiment design skills as

measured by tests improved significantly compared to those of the control group. However, the same approach did not prove to be successful by the end of the first school year of a longitudinal study in case of 12–13-years-old students (Szalay et al., 2020). Accordingly, from the second year onwards, pupils in the experimental groups were taught the relevant principles of experimental design. This scaffolding produced positive results for 13–14 year-old students, since the experimental groups' experimental design skills seemed to develop more than that of the control group's. However, this effect disappeared in the long term when the students turned 14–15 (Szalay et al., 2021). This led to the conclusion that students probably need more help and more motivation in designing experiments than they received in the previous longitudinal project. Gott and Dugan (1998) warned that not all inquiry based laboratory tasks are appropriate to engage students in scientific practices, as they depend on their structure and requirements. This is in agreement with Baird's view (1990) that purposeful inquiry does not happen spontaneously – it must be learned. Students obviously need scaffolding to solve inquiry type tasks (e.g. Puntambekar and Kolodoner, 2005; Blanchard et al., 2010; Crujeiras-Pe´rez and Jime´nez-Aleixandre, 2017). This might help to alleviate the increased cognitive load.

Therefore, in the present four-year project that began in September 2021, pupils in the experimental groups answer a series of questions concerned with the design of the experiment about the fair testing (t.e. changing one factor at a time while keeping all other conditions the same) on their worksheets. This is a simplified version of the Experiment Design Diagram described by Cothron et al. (2000). After the first school year of the present project, it was clear that the applied type of instruction had a significant positive effect on the results of the pupils who were required to design the experiments, guided by that set of questions (Szalay et al., 2023).To increase motivation in the present longitudinal research project, context-based tasks with elements of systems thinking were also introduced in the worksheets under the heading "Let's think!". These are the same for all groups.

Aims and Objectives

Since the method used in the first year of the present four-year longitudinal research (Szalay et al., 2023) seemed to improve the experimental design skills of the experimental group (Group 3) who had to answer questions about the design of the experiments before they planned the steps of the experiments they carried out, it was decided to apply the same research model in the following years to see what changes happen in the longer term. It is also interesting to see how the performance of the other experimental group (Group 2) who answer the questions after completing the step-by-step experiments changes over the course of the tests.

Research Questions (RQ)

Therefore, in the second school year of the present project, answers to the same research questions as in the previous year were sought.

RQ1: Did the intervention result in a significant change in pupils' ability to design experiments (experiment design skills, EDS) in either of the experimental groups compared to the control group in long term, by the end of the second year of the present project?

RQ2: Did the pupils in the experimental groups score significantly differently on the disciplinary content knowledge (DCK) questions because of the intervention compared to the students in the control group in long term, by the end of the second year of the present project?

RQ3: Was there a difference in EDS between students in the two experimental groups in long term, by the end of the second year of the present project?

Methods

Research Design and Participants

A quasi-experimental design with a non-equivalent control group is applied in this empirical research. The research team consisted of thirty-four serving chemistry teachers and five university chemistry lecturers at the beginning of this four-year project (in September 2021). Thirty-one of the teachers taught the participating students in the first year of the project. Three teachers did not teach the students in the sample. One of them, as a member of the research team, tried out the tests with her pupils and prepared worksheets. Another teacher is involved in correcting the marking of the tests that had been done by other teachers. A third teacher only prepared one of the worksheets. Of the thirty-four in-service chemistry teachers, five teachers have left the research team since the end of the first year of this project. Two teachers from the participating schools who replaced two of the five teachers who left became members of the research team. Therefore, the research team now consists of thirty-one serving chemistry teachers and five university chemistry lecturers. All teachers are voluntary participants.

Participating pupils must attend a school where they are taught chemistry from Grade 7 to Grade 10 (from age 12–13 to age 16–17), so that their learning of chemistry over four school years can be followed in the present longitudinal research. The 931 seventh-grade pupils who were involved in the beginning of this project (in September 2021) came from twenty-five Hungarian secondary schools and thirty-eight classes. Class sizes varied between 14 and 36, reflecting the typical class sizes in Hungarian schools. The students who remained in the project in the second year came from twenty-three Hungarian secondary schools and thirty-six classes. Class sizes varied between 13 and 33 in the second school year.

At the beginning of the project (September-October 2021), 931 participating seventh-grade pupils completed one test (called Test 0, T0). The 38 classes were grouped into Groups 1, 2 and 3 after the evaluation of the results of Test 0 to ensure that there were no significant differences among them neither in the initial performance (previous knowledge), nor in terms of the hypothesised parameters (school ranking, mother's education, gender). Pupils stay for four years in the same group they were in when the project started. By the end of the first school year (May-June 2022), 890 of these students completed another test (called Test 1, T1). The 756 remaining eighth-graders completed the third test (Test 2, T2) by the end of the second school year (April-May 2023).

Six pupil worksheets and teacher's guides were produced in both school years 2021/22. and 2022/23. Each worksheet was written in three versions for the three groups of students.

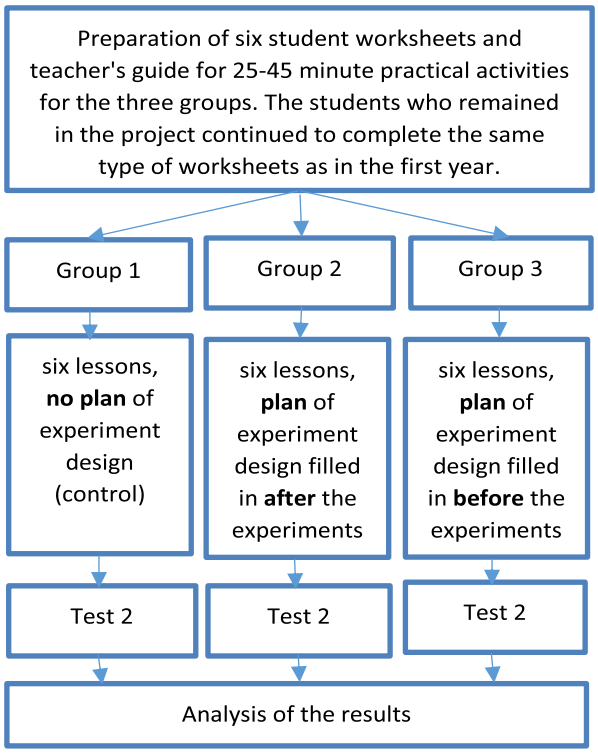
The research model used in the second year of this project is summarised in Figure 1. The teachers chose when the lessons took place, in which the worksheets and the tests provided were used.

Group 1 (the control group) only did step-by-step experiments that Banchi and Bell (2008) called “structured inquiry”. An abbreviated and simplified version of the Experiment Design Diagram (Cothron et al., 2000), applying the ‘fair testing’ method (i.e. changing one factor at a time while keeping all other conditions the same), was used to help experimental groups learn to design experiments. The questions concerned the control of variables, the discussion of hypotheses and the definition of the steps of the experiments. Group 2 carried out the same step-by-step experiments as Group 1, but after performing the experiments, they had to answer questions on the worksheets about the design of the experiments, following the relevant parts of the fair testing plan described above. Their answers were discussed with their teachers. This instruction method could be considered as a much-simplified version of the approach used by Reynders et al. (2019) who linked the discussion questions of the experimental procedure to the lab report. Group 3 were given guided inquiry tasks on their worksheets, since it is more realistic to introduce guided inquiry (Schoffstall and Gaddis, 2007) practicals at school level, where the question to be investigated is given by the teacher and pupils have to design an experiment to find the answer than the open inquiry that is often considered too

challenging even at undergraduate level (Farley et al., 2021). Group 3 students' experiment design was supported by questions from the fair-testing plan mentioned above. The answers were discussed with Group 3 by their teachers before the experiments were carried out. The treatment of Group 3 is similar to that of the "critical thinking" pre-lab group described by van Brederode et al. (2020), as they had to design the experiment to be performed. The "paved-road" pre-lab group of the same study (van Brederode et al., 2020) on the other hand, was given compact lab instructions to carry out the experiment, similarly to Group 2. (Although Group 2 had to discuss aspects of the experiment design after the experiment was completed.)

Figure 1

Research procedure applied in the second school year of the current project



Ethical Considerations

In the absence of institutional ethics committees or local procedures, our research team had to develop its own ethical protocol to ensure that informed consent was obtained and that the privacy and confidentiality of the individuals were protected (Lawrie et al., 2021). A letter describing the key features of the research was drafted in accordance with the General Data Protection Regulation (GDPR) in the European Union and sent to the mother or guardian of each participating student by their chemistry teachers to ask for written consent for their children to participate. Teachers also had written permission from school principals to participate. Teachers told the pupils that the test results would not count in their school's chemistry assessment, but that they were participating in a project to improve chemistry education.

Instrumentation

Worksheets

The pupil worksheets and teacher's guides describe practical activities involving pupil experiments, designed to take about 25-45 minutes. All twelve worksheets and their teacher's notes titled “Student sheets 1-6 and their teacher's notes” for the first year of the present project and “Student sheets 7-12 and their teacher's notes” for the second year of the present project are available in English on the research team's website (https://ttomc.elte.hu/publications/92). These were piloted with students working in small teams.

Topics of the experiments and the context-based systems thinking tasks ("Let's think!") were related to the National Core Curriculum of Hungary (2020), together with the experiment design tasks given to Group 3 on the worksheets (Szalay et al. 2023) and see Table 1 in Appendix for those used in the second year.

Each topic of the worksheets had been agreed by all participating teachers. The first versions of each worksheet were read by four university lecturers (who are experts in the development of chemistry teaching materials for primary and secondary school learners. The worksheets were then improved by the authors based on the experts' suggestions. This second version was proofread by one of the experts and the leader of the research team, who then agreed on the final changes.

Each worksheet continued to include a context-based task under the heading ‘Let's think!’, designed to maintain interest, engagement and develop systems thinking skills (e.g. Chen, Xiao, 2020; del Mar López-Fernández et al., 2022; Klemeš, et al., 2021; MacDonald et al., 2022).

The basics of the correct terminology (independent and dependent variables, constants, hypotheses) were introduced for both experimental groups (Groups 2 and 3) in the second school year, in the hope that students would be old enough to master these abstract concepts. Based on answers given to attitude questions in the first year, other minor changes were made to the worksheets, highlighting the importance of experimentation in science. Groups 2 and 3 teachers were also asked to encourage their pupils to answer questions about experimental design by highlighting its usefulness and praising them for thinking well.

Tests

Based on the results of the end-of-year test in the first (Test 1, i.e. T1) and second (Test 2, i.e. T2) year of the project, pupils' achievement should be compared with the performance of the previous year (Test 0, i.e. T0): a test at the beginning of the project and Test 1, (i.e. T1): end-of-year test in the first year of the project, respectively) to see how their DCK and EDS have changed. Tests were designed using the same recommendations as described in the previous studies (National Research Council, 2001; Szalay et al., 2023) and they were all paper based (Cannady et al., 2019). The tasks had to integrate content knowledge that learners are familiar with (DCK) and focus on the ability to apply scientific practices (Zimmerman, 2000; Zimmerman, 2007; OECD, 2017; Cannady et al., 2019; Tosun, 2019) while solving the experimental design tasks (EDS). EDS tasks had to provide the content knowledge needed to solve the tasks (Cannady et al., 2019).

Before designing the EDS task, various assessment criteria (Sirum & Humburg, 2011) assessment tools (Chen et al., 2019; Tseng et al., 2022) and the experimental design checklist for Science Olympiad (2020) were consulted for guidance. The collection and use of data from the text, the identification of independent, dependent and controlled variables, and the procedure were considered most important. The EDS tasks were set in the context of everyday life in a way that was designed to engage the students' interest.

The following EDS task was used in Test 2 to compare the development of pupils' EDS across the three groups.

"Imagine you are on holiday with relatives in a small village and you have filled a swimming pool in the yard with water. The water needs to be disinfected with a tablet that can work in the pH=7-8 range, but the pH of the local water is higher than 8. There is a chemical that needs to be added to the pool in such quantities that the bathing water reaches the desired pH range. However, the test strip to check the pH of the water is out of stock and is only available in a remote town. Remember, however, that red cabbage juice can act as an acid-base indicator and can be made from cabbage at home. According to the Internet, red cabbage juice is yellow at pH ≥12, green and greenish blue around pH=9-11, blue between pH=7-8, lilac or purple between pH=4-7, and red in the pH≤3 range. This allows you to adjust the pH of the pool water to the pH range that the disinfectant tablets will work. You cannot pour the cabbage juice into the pool, but you can use cups to take a water sample from the pool, even several times. There is also a shovel to mix the water in the pool. Use your answers below to help design the experiments to get the right pH range.

a) What can you change about the total content of the pool during the experiments (i.e. what should be added to the total content of the pool in each experiment)?

b) What property of the pool water depends on the change you cause?

c) How can you test for this property of the pool water mentioned in b) above?

d) From what observation can you conclude that more material needs to be added?

e) Why is it always important to mix the pool water carefully?

f) In which case can you conclude that you can now put the disinfectant tablets in the pool water?

g) Put a (+) sign in front of the statement(s) in the list below that is/are important and a (-) sign in front of the statement(s) that is/are not important. (You can write a different sign after a clear strike-through if you change your mind.)

The water should be always taken out from the same point in the pool by the cup.

Always the same volume of water should be taken out from the pool by the cup.

Always the same cup should be used to take out the water from the pool."

The test questions were structured according to the levels of the revised Bloom's Taxonomy (Bloom et al., 1956; Krathwohl, 2002) cognitive process dimension categories as interpreted in previous publications (Szalay et al., 2020; Szalay et al., 2021; Szalay et al., 2023). Each test consisted of eighteen compulsory items, each worth 1 point. Nine were used to assess EDS and the other nine to assess DCK (three each for recall, understanding and application), as both experiment design skills (EDS) as part of inquiry skills and disciplinary content knowledge (DCK) had to be assessed (e.g., Cooper, 2013; Reed, Holme, 2014; Rodriguez, Towns, 2018; Underwood et al., 2018). To measure the development of the experimental design skills (EDS) problem solving tasks were used that required the application of the components of experimental design skills defined by Csíkos et al. in 2016 (i.e. identification and control of variables, including the principle of "how to vary one thing at a time" or "other things/variables held constant"; choosing equipment and materials).

The instruments used were 40 minutes paper-based tests. The papers were coded so that teachers would know the the respondent's name and gender, but the researchers only received anonymous data coded for statistical analysis. These codes and the Excel spreadsheets containing the codes and marks are used throughout the project. Participating teachers marked the tests, recording the marks in an Excel spreadsheet as instructed (see on the research group website (https://ttomc.elte.hu/publications/95) under the titles "T0 test and instructions for teachers", "T1 test and instructions for teachers" and "T2 test and instructions for teachers", respectively). As there was an element of subjectivity in the marking protocol, the research group tried to standardise the marking to ensure that the application of the marking key is the same for each corrected test of the same kind, as done by Goodey & Talgar (2016). An experienced chemistry teacher reviewed all the teachers' marking and suggested modifications to the marking instructions. After discussions within the team, alterations were made. Based on these, the teachers' marks were changed to ensure that a unified marking process, free from individual teachers' decisions was used. The scoring procedure is consistent with the recommendation of reaching complete consensus through negotiated agreement (Watts, Finkenstaedt-Quinn, 2021).

Validity

Evidence for content validity was established by a panel of domain experts judging whether the items appropriately sample the domain of interest (Crocker & Algina, 2006). It can be used to argue against construct underrepresentation that is one of the main threats to construct validity (Wren & Barbera, 2013). To avoid construct-irrelevant variance, each task of Test 2 could be completed after finishing the tasks on the six worksheets for the second school year of the present project. Table 2 shows how each task of Test 2 can be matched with the relevant content of certain worksheet.

In Test 2, the tasks had to be different than in previous tests to avoid repeated testing effects and to measure transferable EDS (Cannady et al, 2019; Schafer et al., 2023; Szalay et al, 2020; Szalay et al, 2021; Szalay et al, 2023). The chances of the successful solution of a task would be higher if it was used the second time, since pupils might discuss it with others in between times. (This could have caused construct-irrelevant easiness.)

Table 2

Matching the content of the tasks of t2 test and the topic(s) of the worksheets for the second year

No. of task in Test 2	No. of student worksheet and topic
1.a-b	8. Hardness of water, use of water softeners (precipitation reactions)
2. a-g	10. Modelling environmental processes (acid base reactions, pH, effects of acid rain)
3. a-b	9. Modelling industrial processes (production and use of quicklime and slaked lime)
4.a-b	7. Reactivity series of metals and hydrogen (redox reaction, electron transfer)
5.	12. Plastics - pros and cons (raising environ-mental awareness)
6.	9. Modelling industrial processes (production and use of quicklime and slaked lime)
7.	11. Modelling qualitative analysis (health and diet)

The first version of Tests 2 and its marking key was devised by the research team leader. Then the same university educators in the research group who checked the content of the student worksheets and the T0 and T1 tests checked the marking instructions of the T2 test. Alterations were made according to their suggestions. This process of item evaluation and revision took place for all items of all the three tests. Expert feedback on item content, wording, and consensus of the correct answer are all sources for evidence of expert response process validity and against construct-irrelevant variance, both construct irrelevant difficulty and easiness (Wren & Barbera, 2013).

Test 2 was trialled with two classes (N1=29, N2=29, altogether 58) of 13 year-old pupils not participating in the research in the autumn 2022. (Test 0 and 1 had been tried out on the same two classes in the previous school year). The chemistry teacher organising this pilot and marking each test gave detailed suggestions how to improve the wording of the tasks and the marking instructions based on her experiences. T2 test and their marking instructions were further revised in response to results of the trial before they were filled in by the pupils participating in the sample.

Participating teachers had not seen Test 2 before piloting the six student worksheets of the school year. The aim was to ensure that the tasks in Test 2 did not subconsciously influence teachers' teaching behaviour, which could have affected pupils' responses to the test questions.

The test scores of Groups 2 and 3 were compared with those of Group 1 (control group) to eliminate the risk of maturation (Shadish et al., 2002).

Data Collection

The number (N) completing all three tests (T0, T1 and T2) in each group is as follows: Group 1: 242; Group 2: 273; Group 3: 241, altogether 756. Following the incompletion of a test, that pupil was excluded from the analysis and future tests. Further, two entire classes no longer participated in the second year of the research because their teachers resigned, and the new teachers did not volunteer to continue working in our research team.

The following data were collected and analysed statistically:

- Total scores for Test 0, Test 1 and Test 2.
- Scores for EDS tasks Test 0, Test 1 and Test 2.
- Scores for DCK tasks Test 0, Test 1 and Test 2.
- Gender.
- School ranking. The student's school ranking amongst Hungarian secondary schools, according to the website 'legjobbiskola.hu'. The participating schools were grouped into high, medium, and low-ranking categories and a categorical variable was used according to these three levels. This allowed a statistical assessment of the impact of participating schools "quality" on the development of the pupils' knowledge and skills.
- Mother's education. Two categories were formed depending on whether or not the student's mother (or guardian) had a degree in higher education. This categorical variable was intended to characterise the student's socioeconomic status.

Statistical Methods

In constructing the three groups, care was taken to ensure that they did not differ in terms of the previous knowledge (measured by Test 0) and neither of the hypothesised parameters (school ranking, mother's education, gender). This was checked by a chi-square test.

Cronbach's alpha values (Cronbach & Meehl, 1955) for the three tests were acceptable: 0.742 for T0 test, 0.692 for T1 test and 0.694 for T2 test.

Statistical analysis of data was done by the SPSS Statistics software. ANOVA and ANCOVA analyses were also performed. According to Howell (2012), ANCOVA can be used to adjust for the initial difference and to reflect the effect on the dependent variable. Raw mean scores (before ANCOVA analysis) and their standard deviations (SD) for the three groups were calculated for all the three tests (T0, T1 and T2) in the whole test ('TOTAL'), the DCK tasks and the EDS tasks. The effect of the intervention on the development of the experimental groups (Group 2 and Group 3) was shown by the Cohen's d effect size (Cohen, 1988). The Cohen's d effect size values were calculated taking into consideration the means and standard deviations of the three types of difference between the three test scores (T1 – T0, T2 – T1 and T2 – T0).

Although the Cohen's d effect size can be used to characterise the effect of development, it was assumed that apart from the three types of instructional methods used during the intervention for Group 1, 2 and 3, other hypothesised parameters (school ranking, mother's education, gender) and a covariate (prior knowledge, i.e. student scores for T0 test) had also influenced the results. Therefore, the statistical analysis of data was also accomplished by analysis of covariance (ANCOVA) to examine the effect in more detail. Effect sizes in the ANCOVA analysis were characterised by the calculated Partial Eta Squared (PES) values. In the case of multiple comparisons Bonferroni correction was applied. While testing the differences among groups and sub-groups, a significance value of $p<0.05$ was applied. However, a significance value of $p<0.025$ was used in the comparison of the results of Test 0 and Test 1, Test 1 and Test 2, Test 0 and Test 2, respectively (according to the Bonferroni correction).

It has also been considered that the results may be biased by the number of chemistry lessons per week that the groups of students in the sample have. Therefore, an ANCOVA analysis was

conducted in which the sum of the number of chemistry lessons per week was also a covariate. However, the resulting PES values varied between 0.000-0.003 and were not significant.

Findings

According to the chi-squared test, there is no significant difference in the composition of the groups with respect to mother's education [X^2 (2, N = 756) = 2.844, p = 0.241)] and gender [X^2 (2, N = 756) = 2.523, p = 0.283)]. However, there is a significant difference in the composition of the groups with respect to school ranking [X^2 (4, N = 756) =13.86, p = 0.008)], as the difference is significant between Group 1 and Group 3 [X^2 (2, N = 517) =11.81, p = 0.003)]. This may be mainly due to the fact that Group 3 has a higher proportion in high ranking schools and a lower proportion of pupils in medium ranking schools than the other two groups.

Table 3 shows the raw mean scores, prior to ANCOVA analysis, and their standard deviations (SD) for the three groups for the T0 test for the whole test ("TOTAL"), the DCK tasks ("DCK"), the EDS tasks ("EDS") and the results of the ANOVA analysis. High standard deviations show that the sample was very heterogeneous according to their knowledge and skills as measured by the tests. ANOVA analysis revealed no significant difference between groups in the performance of either T0TOTAL or T0DCK or T0EDS.

Table 3

The means of scores and their sd-s for the whole test ("total"), the dck tasks and the eds tasks of t0 and the results of the anova analysis (n=756)

Group	T0TOTAL (SD)*	T0DCK (SD)**	T0EDS (SD)**
Group 1	11.39 (3.76)	5.57 (1.77)	5.82 (2.64)
Group 2	11.47 (3.21)	5.65 (1.76)	5.82 (2.40)
Group 3	11.04 (3.55)	5.54 (1.69)	5.51 (2.57)
$F(2, N = 756)$	1.05	0.287	1.25
p	0.350	0.750	0.286
Sign.	-	-	-

Note: *Maximum scores: 18; **: Maximum scores: 9

Table 4 shows the mean raw scores, their standard deviations and the results of the ANOVA analysis for the three groups for T1 test. In all cases, the average raw scores for T1 were lower than for T0 tasks. This is understandable, as the knowledge and skills measured by T1 (and T2) were different to those measured by T0. (The three tests contained different tasks for the reasons explained earlier under the heading "Validity"). There is a significant difference among the performance of groups in T1TOTAL, T1DCK and T1EDS. The achievement of Group 3 exceeded that of the other two groups in the end of the first year (Grade 7) of this project., whereas Group 2 performed worse in T1 (because they scored significantly lower in T1DCK) than the control group (Group 1) and the other experimental group (Group 3). These results are consistent with previously published trends (Szalay et al., 2023).

Table 4

The means of scores and their sd-s for the whole test ("total"), the dck tasks and the eds tasks of t1 and the results of the anova analysis (n=756)

Group	$T1_{TOTAL}$ (SD)*	$T1_{DCK}$ (SD)**	$T1_{EDS}$ (SD)**
Group 1	9.16 (3.66)	4.52 (2.14)	4.64 (2.15)
Group 2	8.55 (3.15)	4.04 (1.71)	4.50 (2.08)
Group 3	9.94 (3.40)	4.55 (2.00)	5.39 (2.11)
$F(2, N = 756)$	10.75	5.50	12.68
p	0.000	0.004	0.000
Sign.	2 < 1 < 3	2 < 1, 3	1, 2 < 3

Note: *Maximum scores: 18; **: Maximum scores: 9

Table 5 shows the mean raw scores and their standard deviations for the three groups for T2 test. Group 3 again outperformed the other two groups at the end of the second year of the project (Grade 8) in the experimental design tasks ($T2_{EDS}$) and thus in the whole test ($T2_{TOTAL}$). There was no significant difference between Group 2 and Group 1 in the results of the T2 test ($T2_{TOTAL}$) and its sub-tests ($T2_{DCK}$ and $T2_{EDS}$).

Table 5

the means of scores and their sd-s for the whole test ("total"), the dck tasks and the eds tasks of t2 and the results of the anova analysis (n=756)

Group	$T2_{TOTAL}$ (SD)*	$T2_{DCK}$ (SD)**	$T2_{EDS}$ (SD)**
Group 1	9.12 (3.36)	3.33 (2.05)	5.79 (2.03)
Group 2	9.53 (3.06)	3.60 (1.95)	5.92 (1.89)
Group 3	9.79 (3.17)	3.64 (2.18)	6.15 (1.85)
$F(2, N = 756)$	2.73	1.61	2.27
p	0.066	0.201	0.104
Sign.	1 < 3	-	1 < 3

Note: *Maximum scores: 18; **: Maximum scores: 9

For further analysis, the dependent variable was the difference between the three test scores (T1 – T0; T2 – T1 and T2 – T0). Based on the means and standard deviations of the differences between the three test scores (T1 – T0; T2 – T1 and T2 – T0), Cohen's d effect size values were also calculated that are presented in Table 15 in Appendix. These results clearly show that Group 2 developed better than the other two groups, especially in DCK tasks in the second year of the project. However, when comparing the results of the three groups over the first two years, the change in performance of Group 3 was still significantly better than the performance in the other two groups in the EDS tasks.

Previous experience had shown that performance can depend on several factors, not only on the intervention. Therefore, an ANCOVA analysis was conducted with test scores as the dependent variable. Group (the type of instruction methods), school ranking, mother's education and gender were the parameters. The covariate was prior knowledge (T0 test scores). This was also necessary because, after the two classes were omitted from the project in the second year, there was a significant difference in the composition of the groups in terms of school ranking. This adjustment can clarify the treatment effect in a research study. Partial Eta Squared (PES) values characterising the effect sizes are shown in Table 6-8.

Initially, as published earlier (Szalay et al., 2023) it was mainly the school ranking and, to a lesser extent, in the DCK tasks, the mother's education that had a significant effect on the scores. After the intervention in the first year (in the T1 test), three parameters were found to have significant effect sizes (PES) on the changes for the whole test and both sub-tests: group, school ranking and prior knowledge. Of these, prior knowledge had the largest effect size on the whole test (Table 6) and in the EDS tasks (Table 8). School ranking, however, had more effect in the DCK tasks than in the EDS tasks,

while the instruction methods ("Group") appeared to have more effect on performance in the EDS tasks than in the DCK tasks.

In the end of the second school year (in the T2 test), the same three parameters still seem to be important in the whole test (Table 6). However, only prior knowledge had a significant effect on changes in DCK tasks (Table 7). As for the changes in the EDS tasks in the second year, mother's education showed a significant but small effect. Among the other three parameters, school ranking had the largest and instruction methods ("Group") had the smallest effect size (Table 8).

Table 6

The effects of the assumed parameters (sources) and the covariate (prior knowledge, t0$_{total}$) on the changes for the whole test ("total") in the beginning of the project (t0), in the end of grade 7 (t1) and in the end of grade 8 (t2) (n=756)

Parameter (Source)	PES (Partial Eta Squared)		
	T0$_{TOTAL}$	T1$_{TOTAL}$	T2$_{TOTAL}$
Group	0.005	0.042*	0.012*
School ranking	0.109*	0.046*	0.009*
Mother's education	0.010*	0.004	0.001
Gender	0.006	0.000	0.000
Prior knowledge (T0$_{TOTAL}$)	-	0.136*	0.102*

Note: * Significant at p < 0.025 level (Bonferroni correction)

Table 7

The effects of the assumed parameters (sources) and the covariate (prior knowledge, t0$_{dck}$) on the changes for the dck tasks ("dck") in the beginning of the project (t0), in the end of grade 7 (t1) and in the end of grade 8 (t2) (n=756)

Parameter (Source)	PES (Partial Eta Squared)		
	T0$_{DCK}$	T1$_{DCK}$	T2$_{DCK}$
Group	0.001	0.018*	0.004
School ranking	0.033*	0.079*	0.001
Mother's education	0.021*	0.002	0.002
Gender	0.009*	0.000	0.001
Prior knowledge (T0$_{DCK}$)	-	0.053*	0.049*

Note: * Significant at p < 0.025 level (Bonferroni correction)

Table 8

The effects of the assumed parameters (sources) and the covariate (prior knowledge, t0 $_{eds}$) on the changes for the eds tasks ("eds") in the beginning of the project (t0), in the end of grade 7 (t1) and in the end of grade 8 (t2) (n=756)

Parameter (Source)	PES (Partial Eta Squared)		
	T0$_{EDS}$	T1$_{EDS}$	T2$_{EDS}$
Group	0.006	0.040*	0.011*
School ranking	0.113*	0.023*	0.059*
Mother's education	0.001	0.005	0.010*
Gender	0.001	0.001	0.002
Prior knowledge (T0$_{EDS}$)	-	0.070*	0.045*

Note: * Significant at p < 0.025 level (Bonferroni correction)

The effects of the assumed parameters "Group" and "School ranking" estimated by the model of the ANCOVA analysis (absolute mean scores) for the whole test, the DCK tasks and the EDS tasks, as well as the significance of their differences for the three tests are shown in the Tables 16-17 in

Appendix. These data show that there is no significant difference among the achievement of the three groups in the DCK sub-test of T2 (Table 16). However, Group 3 significantly outperformed the control group in the EDS tasks in the first two years.

The relative estimated average scores (ratios of the estimated mean scores of the experimental groups compared to that of the control group's) for the whole test and for the sub-tests in the beginning of the project (Grade 7, T0) are shown in Table 9, in the end of first school year (Grade 7, T1) in Table 10 and in the end of second school year (Grade 8, T2) in Table 11.

Table 9

The estimated mean scores of the experimental groups divided by the estimated mean scores of the control group for the whole test ("total"), in the dck tasks ("dck") and eds tasks ("eds") in test 0 (n=756)

Ratio	T0TOTAL	T0DCK	T0EDS
Group 2 / Group 1	1.00	1.01	1.00
Group 3 / Group 1	0.96	0.99	0.93

The data in Table 10 show that, taking the changes in DCK tasks into account, Group 2 performed poorly at the end of the first year compared with the other two groups (Szalay et al., 2023).

Table 10

The estimated mean scores of the experimental groups divided by the estimated mean scores of the control group for the whole test ("total"), in the dck tasks ("dck") and eds tasks ("eds") in test 1 (n=756)

Ratio	T1TOTAL	T1DCK	T1EDS
Group 2 / Group 1	0.93	0.89	0.97
Group 3 / Group 1	1.10	1.00	1.17

However, at the end of the second year (Table 11), the ratio of DCK scores was almost the same in both experimental groups. A significant increase in the performance in the EDS tasks was observed in Group 3 at the end of the first year (Table 10, Szalay et al., 2023). At the end of the second year, both experimental groups performed better in the EDS tasks than the control group, but Group 3 still achieved better results than Group 2 (Table 11).

Table 11

The estimated mean scores of the experimental groups divided by the estimated mean scores of the control group for the whole test ("total"), in the dck tasks ("dck") and eds tasks ("eds") in test 2 (n=756)

Ratio	T2TOTAL	T2DCK	T2EDS
Group 2 / Group 1	1.04	1.07	1.03
Group 3 / Group 1	1.09	1.08	1.08

Next, an ANCOVA analysis was conducted with the changes in test scores (T1 – T0, T2 – T1, T2 – T0) as the dependent variables, group (instruction methods), school ranking, mother's education, and student's gender as the parameters, and the student's prior knowledge (T0 test scores) as the covariate. The results of that ANCOVA analysis are presented in Table 18 in Appendix. These data also show that, among the assumed parameters, mostly the group (type of instruction methods), school ranking and prior knowledge had a significant effect on pupils' performance on the tests. The values estimated by the ANCOVA model showing the effect of the assumed parameters on changes in their performance in the whole tests and sub-tests are shown in the Tables 19-22 in Appendix. Table 19 in Appendix shows that there was no significant difference among the three groups in the development of DCK tasks in the first two years of the project. However, Group 3 performed significantly better in

the EDS tasks than the control group during this period. In the second year of this project, school ranking had a significant effect on scores (Appendix, Table 20). It is interesting to note, however, that the higher the school rank, the lower the change in performance in terms of scores in the DCK tasks in the second year, while in terms of changes in scores in the EDS tasks, students from low-ranking schools performed significantly worse than students from medium- and high-ranking schools. It might also be noteworthy that students with a graduate mother showed weaker progress in the second year in the DCK sub-test than the others, but better progress in the EDS sub-test during the two years (Appendix, Table 21). No significant difference was found between the changes in boys' and girls' performance at any time or in any type of test scores (Appendix, Table 22).

The Cohen's d effect size values calculated by the ANCOVA model from the estimated changes in students' performance in the tests are presented in Table 12 for the whole test, Table 13 for the DCK tasks and Table 14 for the EDS tasks.

Table 12

Cohen's d effect size values calculated by the ancova model from the estimated changes in students' performance on the tests for the whole test ("total") (n=756)

Cohen's d	$T1_{TOTAL} - T0_{TOTAL}$	$T2_{TOTAL} - T1_{TOTAL}$	$T2_{TOTAL} - T0_{TOTAL}$
Group 2 / Group 1	-0.19	0.27	0.12
Group 3 / Group 1	0.25	-0.01	0.24
Group 3 / Group 2	0.44	-0.28	0.12

Table 13

Cohen's d effect size values calculated by the ancova model from the estimated changes in students' performance on the tests for the dck tasks ("dck") (n=756)

Cohen's d	$T1_{DCK} - T0_{DCK}$	$T2_{DCK} - T1_{DCK}$	$T2_{DCK} - T0_{DCK}$
Group 2 / Group 1	-0.24	0.28	0.10
Group 3 / Group 1	0.01	0.10	0.12
Group 3 / Group 2	0.24	-0.18	0.02

Table 14

Cohen's d effect size values calculated by the ancova model from the estimated changes in students' performance on the tests for the eds tasks ("eds") (n=756)

Cohen's d	$T1_{EDS} - T0_{EDS}$	$T2_{EDS} - T1_{EDS}$	$T2_{EDS} - T0_{EDS}$
Group 2 / Group 1	-0.07	0.13	0.09
Group 3 / Group 1	0.34	-0.12	0.23
Group 3 / Group 2	0.41	-0.26	0.14

Discussion

Based on these data, it appears that Group 2 had caught up with Group 3 in the second year in terms of a positive, but quite small, change in performance in DCK tasks compared to that of the control group's (Table 13). Group 2 also improved better than the control group in terms of performance in the EDS tasks in the second year (Table 14). However, over the two years, Group 3 of the two experimental groups still developed better in the EDS tasks than Group 2. It appears, therefore, that both using the questions while designing the experiments to be carried out (for Group 3) and answering the questions after carrying out the step-by-step experiments (for Group 2) can help students to achieve better results in the EDS tasks but using the questions to help to design the experiments in practice (Group 3) still produced better results in long term. Changes in EDS were

higher than changes in DCK. These results are in line with Bredderman (1983), who reported that the use of inquiry-based methods had a greater effect on science process than on science content.

These results might have been caused by the different treatments of the two groups. Group 2 students did not have to plan experiments. Those classes had to discuss with their teacher why the experiments were designed as they were (according to the questions on their worksheets). This could be seen as a "theoretical" method for learning about experimental design, which might take longer for students to realise how to apply it in practice, which could have happened by the end of the second year. We can fully support Potier's (2023) claim that the skills needed to succeed with guided inquiry approaches take time to develop. On the other hand, Group 3 had to design experiments, in teams, while they were answering questions helping to learn experimental design. This can be seen as a direct "practical" way of learning experimental design. This may be the reason why it made an impact in the first year and had a significant positive effect on the experimental design skills by the end of the two years period. The treatment of Group 3 is similar to that of the "critical thinking" pre-laboratory group described by van Brederode et al. (2020). On the other hand, the treatment of Group 2 resembles to the "paved road" pre-laboratory group of the same study. The present findings show that the intervention for Group 3 produced better results, as was the case for the "critical thinking" group. These results are also consistent with the study by Tseng et al. (2022), where evaluative reflection on peers' experimental designs improved students' scientific inquiry performance in formulating research questions and designing experimental procedures more than the recognition of variables in designing experimental procedures. The present findings seem to support Matthews' (2018) claim too that learners can gain meaningful insights into the construction of scientific knowledge through processes of inquiry, reasoning and planning, but only if they are properly organised and reflected upon. The findings in connection with Group 3 also support that adequate and appropriate scaffolds should be provided for students coming from a traditional teaching style to successfully complete an investigation task based on inquiry-based learning (Seery et al., 2019a).

The ANCOVA model calculations show that the mother's education had only a weak significant effect on the development of the experimental design skills in this project in the first two years. This seems to contradict the Education and Training Monitor (2020) report, which shows that socio-economic background is a strong predictor of student performance. This can be explained by the fact that the sample of the present study is not representative of the cohort, as these pupils had gone through a very tough selection process when they took the entrance exam to their current school.

School ranking is still an important parameter according to the present results, as it had a significant impact on EDS scores in both years. This is understandable, as Siegler et al. (2010) argue that school is the microsystem that, alongside the family environment, has the strongest influence on youngsters' development. Within this context, the interaction between teachers and learners has a profound influence on pupils' motivation towards chemistry as a subject.

The gender still did not seem to have any significant effect on the achievement in any type of the test scores in the present study. This is in line with the results of other authors, who did not find any significant difference in students' acquisition of science process skills (SPSs) with respect to gender (Ofoegbu, 1984, Walters & Soyibo, 2001; Böyük et al., 2011; Güden & Timur, 2016). However, Tosun's (2019) study revealed that the most important predictive variables on SPS level were gender, grade level and mothers' education level from the examined demographical features. Onukwo (1995) also found a significant gender difference in the levels of SPS.

The means of the T1 test scores estimated by the ANCOVA model at the end of the first year (Szalay et al., 2023) and the means of the T1 test scores estimated by the ANCOVA model at the end of the second year (after the sample composition changed) were compared. The difference was found to be very small, ranging from 0.2 to 3.6%, with an average difference of 2.0%. Thus, it seems that the changes in the composition of the groups were handled well by this analysis.

Conclusion and Implications

Summary of the Results and Answers to the Research Questions

The statistical analysis of the results measured at the end of the second year of the present four-year project showed similar results to those measured at the end of the first year, in the sense that four of the assumed parameters had a significant effect on the Grade 8 pupils' scores in the tasks intending to measure the experimental design skills: the intervention, the school ranking, the prior knowledge and, to a much lesser extent, the mother's education. Of these four, school ranking seemed to have the biggest impact on performance in the second year. Prior knowledge, which had the highest PES value at the end of the first year, still appears to have a larger effect than the intervention. The intervention did not seem to have a significant effect on scores in the tasks measuring disciplinary content knowledge at the end of the second year of the project.

The answers to the research questions are as follows.

RQ1: By the end of the second year, the intervention resulted in a significant positive change in the experimental design skills (EDS) of Group 3 participants compared to the control group (Group 1), as measured by the tests (Cohen's d: 0.23). It is reasonable to assume that this was due to the fact that the Group 3 worksheets included questions to support experimental design. Although the change in the performance of Group 2 in EDS tasks by the end of the second year was also positive compared to that of the control group's (Cohen's d: 0.09), it was not found significant.

RQ2: By the end of the second year of the present project, no significant difference in the change in disciplinary content knowledge (DCK) among the three groups could be measured (Cohen's d for Group 2: 0.10 and Group 3: 0.12, respectively).

RQ3: No statistically significant difference was found between the mean scores of the two experimental groups, considering the extent to which their experimental design skills developed during the first two years of the project (Cohen's d: 0.14).

It should be noted that in the first year, the change in performance of Group 3 on the EDS tasks was significantly better than the change in performance of Group 2 students (Cohen's d: 0.41). In the second year, however, this trend was reversed, and Group 2 improved better than Group 3 (Cohen's d: -0.26).

Currently, the use of the Group 3 worksheets and similar experimental design tasks using the set of questions can be relatively confidently recommended to practising teachers. However, since Group 2 in the second year of the project performed significantly better than the Group 3 in the EDS tasks of the test, it may be useful for colleagues who are reluctant to teach the experimental design directly to try using the worksheets for Group 2 and to have their pupils answer the questions after completing the step-by-step experiments.

Limitations

Compared to the previous year's sample size, the number of participants decreased. This was due to the two missing full classes and many pupils missing from other classes at the time of the T2 test. This is unfortunate, but the reasons mentioned earlier were beyond the control of the research team.

The sample was not representative of the examined cohort of learners (Grade 8, 13-14 years old). Rather, it was representative of higher achieving students, as they were selected by entrance exams to the participating schools. The reason for this is that the pupils have to stay in the same school for the four years of the project. This only allows those from schools that teach chemistry as a separate subject from Grade 7 to Grade 10 to participate.

No single study can evaluate every variable and every theoretical relationship underlying an instructional model (Mack et al., 2019). In addition, the instruments used (40 minutes paper-based tests) could only provide a limited picture of how pupils had benefited from the interventions. As

reading comprehension is key to performance in science (Neri et al., 2021), it may have also influenced the results.

Performance on any assessment is at least partially driven by the learners' motivation for success on the measure and test taking abilities (Cannady et al., 2019). There is a well-documented positive relationship between affective dimensions (and within these, attitudes, motivation and interest) and academic performance in chemistry (e.g. Wang & Lewis, 2022). This is a particular problem when it comes to measuring change in longitudinal studies such as this one, as learners' motivation to learn science often declines as they move from one grade to the next (Schunk et al., 2014; Vedder-Weiss & Fortus, 2011; Vedder-Weiss & Fortus, 2013). Although the research team tried to find contexts that were likely to be of interest to the learners (see Table 1 in Appendix), probably not everyone was equally engaged.

This research will continue for two more years, following the same research model. It is possible that different results or even different trend changes may be observed in the coming years. There are many random events that can affect the final data. Although the relatively large sample size should compensate for most of these, we can never be absolutely sure (Lawrie, 2021).

Implications

The trends in the development of the two experimental groups in the second year of the project were different from those observed in the first year (Szalay et al., 2023). The experimental design skills of Group 2 seemed to improve significantly more than that of the Group 3 in the second year, while the reverse was true in the first year. From the start of the project to the end of the second year, however, still only Group 3 showed significantly more improvement in experimental design skills than the control group. Therefore, the current results still show that it is probably useful to base practical activities on designing experiments by answering questions to help them through the process. This is because significantly more pupils in Group 3 than in Group 1 seemed to have understood how to do a fair test correctly during the two years of the project. The usefulness of an experimental design plan, a simplified version of the one described by Cothron et al. (2000) seemed to be still justified. When no such scaffolding was used, in the first year of the previous longitudinal study, the development of EDS was not detectable by the tests (Szalay et al., 2020). This is in agreement with Baird's view (1990) that purposeful inquiry does not happen spontaneously – it must be learned. It is also interesting to note that providing scaffolding in problem-based learning also had a positive impact on creative thinking even at university level (Ernawati et al., 2023). Similar to the results published by other authors (e.g. Reynders et al., 2019), the results of the first two years of the present project suggest that school learners need further support to understand the skills of cognitive processes and to see how these skills manifest in their laboratory work. This might help to alleviate their cognitive load.

Social variables, prior knowledge and "school effects" (including the teacher's effect), which the literature (e.g. Snook et al., 2009) considers as variables affecting performance, were also found to be important in both years of this research.

Acknowledgements

This work was supported by the Research Programme for Public Education Development of the Hungarian Academy of Sciences under Grant SZKF-6/2021. Many thanks for all the colleagues' and students' work.

Conflicts of Interest

There are no conflicts of interest to declare.

References

Akuma, F. V., & Callaghan, R. (2019). A systematic review characterizing and clarifying intrinsic teaching challenges linked to inquiry-based practical work. *J. Res. Sci. Teach*., *56*(5), 619–648. doi: 10.1002/tea.21516.

Allred, Z. R., Shrode, A. D., Gonzalez, J., Rose, A., Green, A. I., Swamy, U., Matz, R. L., & Underwood, S. M. (2022). Impact of Ocean Acidification on Shelled Organisms: Supporting Integration of Chemistry and Biology Knowledge through Multidisciplinary Activities. *J. Chem. Educ*. *99*(5), 2182−2189. doi: 10.1021/acs.jchemed.1c00981.

Arnold, J. C., Boone, W. J., Kremer, K., & Mayer, J. (2018). Assessment of Competencies in Scientific Inquiry Through the Application of Rasch Measurement Techniques, , *Educ. Sci*., *8*(4), 184–203. doi: 10.3390/educsci8040184.

Arnold, J. C., Mühling, A., & Kremer, K. (2021). Exploring core ideas of procedural understanding in scientific inquiry using educational data mining. *Res. Sci. Technol. Educ*., *41*(1), 1–21., doi: 10.1080/02635143.2021.1909552.

Baird, J. R. (1990). Metacognition, purposeful inquiry and conceptual change, In E. Hegarty-Hazel. (Ed.), *The student laboratory and the science curriculum.* London: Routledge. pp. 183−200.

Banchi, H., & Bell R., (2008). The many levels of inquiry, *Sci. Child.*, *46*(2), 26–29.

Beaumont-Walters, Y., & Soyibo K. (2001). An Analysis of High School Student's Performance on Five Integrated Science Process Skills, *Res. Sci. Techol. Educ*., *19*(2), 133–143. doi: 10.1080/02635140120087687.

Belova, N., & Krause, M. (2023). Inoculating students against science-based manipulation strategies in social media: debunking the concept of 'water with conductivity extract'. *Chem. Educ. Res. Pract.*, *24*(1), 192–202. doi: 10.1039/D2RP00191H.

Blanchard, M. R., Southerland, S. A., Osborne, J. W., Sampson, V. D., Annetta, L. A., & Granger, E. M. (2010). Is inquiry possible in light of accountability?: a quantitative of the relative effectiveness of guided inquiry and verification laboratory instruction, *Sci. Ed., 94*, 577–616. doi: 10.1002/SCE.20390.

Böyük, U., Tanık, N., & Saracogˇlu, S. (2011), Analysis of the scientific process skill levels of secondary school students based on different variables, *J. TUBAV Sci*., *4*(1), 20–30.

Burke, K. A., Greenbowe, T. J., & Hand B. M. (2006). Implementing the Science Writing Heuristic in the Chemistry Laboratory. *J. Chem. Educ*., *83*(7), 1032–1038. doi: 10.1021/ed083p1032.

Cannady, M. A., Vincent-Ruz, P., Chung, J. M., & Schunn, C. D. (2019). Scientific sensemaking supports science content learning across disciplines and instructional contexts. *Contemp. Educ. Psychol*., *59*(10), 101802. doi: 10.1016/j.cedpsych.2019.101802.

Chen, L., & Xiao, S. (2021). Perceptions, challenges and coping strategies of science teachers in teaching socioscientific issues: A systematic review. *Ed. Res. Rev.*, *32*(2), 100377. doi: 10.1016/j.edurev.2020.100377.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences,* 2nd ed. Routledge. doi: 10.4324/9780203771587.

Cooper, M. M. (2013). Chemistry and the Next Generation Science Standards. *J. Chem. Educ*., *90*(6), 679–680.

Cothron, J. H., Giese, R. N., & Rezba, R. J. (2000). *Students and Research: Practical Strategies for Science Classrooms and Competitions*, 3rd ed. Dubuque, IA: Kendall/Hunt Publishing Company.

Crocker, L., & Algina, J. (2006). *Introduction to Classical and Modern Test Theory*, 2nd ed., Wadsworth Publishing Company: Belmont, CA.

Cronbach. L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychol. Bullet.*, *52(4)*, 281−302. doi: 10.1037/h0040957.

Crujeiras-Pérez B., & Jiménez-Aleixandre M. P. (2017). High school students' engagement in planning investigations: findings from a longitudinal study in Spain, *Chem. Educ. Res. Pract*., *18*(1), 99–112. doi: 10.1039/C6RP00185H.

Csíkos, Cs., Korom, E., & Csapó, B. (2016). Tartalmi keretek a kutatásalapú tanulás tudáselemeinek értékeléséhez a természettudományokban. *Iskolakultúra*, *26*(3), 17–29. doi: 10.17543/ISKKULT.2016.3.17.

del Mar López-Fernández, M., González-García, F., & Franco-Mariscal, A. J. (2022). How Can Socio-scientific Issues Help Develop Critical Thinking in Chemistry Education? A Reflection on the Problem of Plastics. *J. Chem. Educ*., *99*(10), 3435–3442. doi: 10.1021/acs.jchemed.2c00223.

Education and Training Monitor (2020), Luxembourg: Publications Office of the European Union. PDF. doi: 10.2766/984100.

Ernawati, M. D. W., Yusnidar, Haryanto, Rini, E.F.S., Aldila, F.T., Haryati,T. & Perdana, R. (2023). Do creative thinking skills in problem-based learning benefit from scaffolding?. *Journal of Turkish Science Education*, *20*(3), 399-417.

European Union. (2016). *Horizon 2020: Work Programme 2016–2017: Science with and for Society. European Commission Decision C(2016)1349 of 9 March 2016*. Retrieved from: h2020-wp1617-swfs_en.pdf (europa.eu).

Farley, E. R., Fringer, V., & Wainman, J. W. (2021). Simple Approach to Incorporating Experimental Design into a General Chemistry Lab. *J. Chem. Educ*., *98*(2)**, 350–356.** doi: 10.1021/acs.jchemed.0c00921.

Goodey, N. M., & Talgar, C. P. (2016). Guided inquiry in a biochemistry laboratory course improves experimental design ability. *Chem. Educ. Res. Pract*., *17*(4), 1127-1144. doi: 10.1039/C6RP00142D.

Gott, R., & Duggan S. (1998). Understanding Scientific Evidence –Why it Matters and How It Can Be Taught, in Ratcliffe M.(ed.), ASE (The Association for Science Education) Guide to Secondary Science Education, Cheltenham: Stanley Thornes, pp. 92–99.

Güden, C., ve Timur, B. (2016). Ortaokul öğrencilerinin bilimsel süreç becerilerinin incelenmesi (Çanakkale örneği) [Examining secondary school students' cognitive process skills (Canakkale sample)], Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi,*16*(1), 163–182.

Hendra, Y. A. (2022). Considering the hexad of learning domains in the laboratory to address the overlooked aspects of chemistry education and fragmentary approach to assessment of student learning. *Chem. Educ. Res. Pract*., *23*(3), 518–530. doi: 10.1039/D1RP00271F.

Hendra, Y. A.., & Seery, M. K. (2017). Reasserting the role of pre-laboratory activities in chemistry education: a proposed framework for their design. *Chem. Educ. Res. Pract*., *18*(4), 518–532. doi: 10.1039/C7RP00140A..

Hennah, N., Newton, S., & Seery, M. K. (2022). A holistic framework for developing purposeful practical work. *Chem. Educ. Res. Pract*., *23*(3), 582–598. doi: 10.1039/D1RP00168J.

Jiménez-Aleixandre, M. P., & Erduran, S. (2007). Argumentation in science education: An overview. In S. Erduran, & M. P. Jiménez-Aleixandre. (Eds.), *Argumentation in Science Education. Perspectives from Classroom-Based Research*. Springer: Dordrecht. (pp. 3–27).

Johnstone, A. H. (1997). Chemistry teaching – Science or alchemy? 1996 Brasted lecture. *J. Chem. Educ., 74*(3), 262–268. doi: 10.1021/ed074p262.

Johnstone, A. H. (2006). Chemical education research in Glasgow in perspective. *Chem. Educ. Res. Pract., 7*(2), 49–63. doi: 10.1039/B5RP90021B

Kirschner, P. A. (1992). Epistemology, practical work and academic skills in science education. *Sci. Educ*., *1*, 273–299. doi: 10.1007/BF00430277.

Klemeš, J. J., Fan, Y. V., & Jiang, P. (2021). Plastics: friends or foes? The circularity and plastic waste footprint. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *43*(13), 1549–1565. doi: 10.1080/15567036.2020.1801906.

Koomson, A., Kwaah, C.Y., & Adu-Yeboah, C. (2024). Effect of science process skills and entry grades on academic scores of student teachers. *Journal of Turkish Science Education*, *21*(1), 118-133. doi: 10.36681/tused.2024.007.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview, *Theory Into Practice, 41*(4), 212–218. doi: 10.1207/s15430421tip4104_2.

Lavonen, J., Ávalos, B., Upadyaya, K., Araneda, S., Juuti, K., Cumsille, P., Inkinen, J., & Salmela-Aro, K. (2021). Upper secondary students' situational interest in physics learning in Finland and Chile, *Int. J. Sci. Ed*., *43*(16), 2577-2596, doi: 10.1080/09500693.2021.1978011.

Lawrie, G. (2021). Considerations of sample size in chemistry education research: numbers do count but context matters more! *Chem. Educ. Res. Pract*., *22*(4), 809–812. doi: 10.1039/D1RP90009A.

Lawrie, G. A., Graulich, N., Kahveci, A., & Lewis, S. E. (2021). Ethical statements: a refresher of the minimum requirements for publication of Chemistry Education Research and Practice articles, *Chem. Educ. Res. Pract*., *22*(2), 234–236. doi: 10.1039/D1RP90003J.

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Rev. Educ. Res*., *86*(3), 681–718. doi: 10.3102/0034654315627366.

Linn, M. C., Davis, E. A., & Bell, P. (2004). Inquiry and technology. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education*. New York: Routledge. pp. 3–28. doi: 10.4324/9781410610393.

MacDonald, R. P., Pattison, A. N., Cornell, S. E., Elgersma, A. K., Greidanus, S. N., Visser, S. N., Hoffman, M., & Mahaffy, P. G. (2022). An Interactive Planetary Boundaries Systems Thinking Learning Tool to Integrate Sustainability into the Chemistry Curriculum. *J. Chem. Educ*., *99*(10), 3530−3539. doi: 10.1021/acs.jchemed.2c00659.

Mack, M. R., Hensen, C., & Barbera, J. (2019). Metrics and Methods Used To Compare Student Performance Data in Chemistry Education Research Articles. *J. Chem. Educ*. *96*(3), 401–413. doi: 10.1021/acs.jchemed.8b00713.

Mahaffy, P. G., Brush, E. J., Haack, J. A., & Ho, F. M. (2018). Journal of Chemical Education Call for Papers Special Issue on Reimagining Chemistry Education: Systems Thinking, and Green and Sustainable Chemistry. *J. Chem. Educ*., *95*(10), 1689–1691. doi: 10.1021/acs.jchemed.8b00764.

Matthews, M. R. (Ed.) (2018). *History, Philosophy and Science Teaching – New Perspectives*, Cham: Springer. doi: 10.1007/978-3-319-62616-1.

Moog, R. S., & Spencer, J. N. (Eds.) (2008). *Process Oriented Guided Inquiry Learning (POGIL);* American Chemical Society, Division of Chemical Education: Washington, DC. Retrieved from: Process Oriented Guided Inquiry Learning (POGIL) (acs.org)

Mostafa, T., Echazarra, A., & Guillou, H. (2018). *The science of teaching science: An exploration of science teaching practices in: PISA 2015*, OECD Education Working Papers Series, No. 188, doi: 10.1787/f5bd9e57-en.

National Core Curriculum of Hungary, (2020), 5/2020. (I. 31.) Korm. rendelet A Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról szóló 110/2012. (VI. 4.) Korm. rendelet módosításáról, Magyar Közlöny, 2020. jan. 31., 17, pp. 290–446. Retrieved from: A Kormány 5/2020. (I. 31.) Korm. rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról szóló 110/2012. (VI. 4.) Korm. rendelet módosításáról – eGov Hírlevél

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

Neri, N. C., Guill, K., & Retelsdorf, J. (2021). Language in science performance: Do good readers perform better? *Eur. J. Psychol. Educ*., *36*(1), 45–61. doi: 10.1007/s10212-019-00453-5.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, by States* (Appendix F – Science and Engineering Practices). In Achieve, Inc. behalf twenty-six states partners that Collab. NGSS, (November). pp. 1–103.

OECD. (2016). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*, PISA, OECD Publishing, Paris, doi: 10.1787/9789264267510-en.

Ofoegbu, L. I. J. (1984). *Acquisition of science process skills among elementary pupils in some northern states of Nigeria* [Unpublished PhD dissertation]. Nsukka: University of Nigeria.

Onukwo G. I. N. (1995). *Development and validation of a test of science process skills in integrated science* [Unpublished PhD dissertation]. University of Nigeria.

Orgill, M., York, S., & MacKellar, J. (2019). Introduction to Systems Thinking for the Chemistry Education Community. *J. Chem. Educ*., *96*(12), 2720−2729. doi: 10.1021/acs.jchemed.9b00169.

Oxford Learner's Dictionaries. (2024). enquiry noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com Accessed: April 11. 2024.

Pedaste, M., Maeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle, *Educ. Res. Rev., 14*, 47–61. doi: 10.1016/j.edurev.2015.02.003.

Potier, D. N. (2023). The Use of Guided Inquiry to Support Student Progress and Engagement in High School Chemistry. *J. Chem. Educ*., *100*(2), 1033–1038.

Puntambekar, S., & Kolodoner, J. L. (2005). Toward implementing distributed scaffolding: helping students learn science from design, *J. of Res. in Sci. Teach., 42*(2), 185–271. doi: 10.1002/tea.20048.

Reed, J. J., & Holme, T. A. (2014). The Role of Non-Content Goals in the Assessment of Chemistry Learning. In: L. K. Kendhammer, & K. L. Murphy (Eds.), *Innovative Uses of Assessment for Teaching and Research*. American Chemical Society: Washington, DC. pp. 147–160. doi:10.1021/BK-2014-1182.CH009

Reynders, G., Suh, E., Cole, R. S., & Sansom, R. L. (2019). Developing Student Process Skills in a General Chemistry Laboratory. *J. Chem. Educ*., *96*(10), 2109−2119. doi: 10.1021/acs.jchemed.9b00441.

Rocard, M. (2007). *Science Education NOW: A Renewed Pedagogy for the Future of Europe*. Brussels: European Commision. Directorate-General for Research, rapportrocardfinal.pdf (europa.eu). Accessed: April 11. 2024.

Rodriguez, J. M. G., & Towns, M. H. (2018). Modifying Laboratory Experiments to Promote Engagement in Critical Thinking by Reframing Prelab and Postlab Questions. *J. Chem. Educ*., *95*(12), 2141–2147. doi:10.1021/ACS.JCHEMED.8B00683.

Russ, R. S. (2014). Epistemology of science vs. epistemology for science. *Sci. Educ*., *98*(3), 388–396. doi: 10.1002/sce.21106.

Schafer, A. G. L., Kuborn, T. M., Cara, E., Schwarz, C. E., Deshaye, M. Y., & Stowe, R. L. (2023). Messages about valued knowledge products and processes embedded within a suite of transformed high school chemistry curricular materials. *Chem. Educ. Res. Pract*., *24*(1), 71–88. DOI:10.1039/d2rp00124a

Schoffstall, A. M., & Gaddis, B. A. (2007). Incorporating Guided-Inquiry Learning into the Organic Chemistry Laboratory. *J. Chem. Educ., 84*(5), 848. doi: 10.1021/ed084p848

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*; Houghton Mifflin and Company.

Siegler, R. S., Deloache, J. S., & Eisenberg, N. (2010). *Study Guide for How Children Develop*. Worth Publishers

Snook, I., Clark, J., Harker, R., O'Neil, A.-M., & O'Neil, J. G.. (2009). Invisible Learnings: A commentary on John Hattie's book Visible learning: A synthesis of over 800 metaanalyses relating to achievement, *New Zealand Journal of Educational Studies, 44*(1), 93–106.

Szalay L., & Tóth Z. (2016), An inquiry-based approach of traditional 'step-by-step' experiments. *Chem. Educ. Res. Pract., 17*(4), 923–961. doi: 10.1039/C6RP00044D.

Szalay L., Tóth Z., & Borbás, R. (2021). Teaching of experimental design skills: results from a longitudinal study. *Chem. Educ. Res. Pract*., ***22*(4)**, 1054–1073. doi: 10.1039/D0RP00338G.

Szalay L., Tóth Z., & Kiss, E. (2020). Introducing students to experimental design skills. *Chem. Educ. Res. Pract*., ***21*(1)**, 331–356. doi: 10.1039/C9RP00234K.

Szalay, L., Tóth, Z., Borbás, R., & Füzesi, I. (2023). Scaffolding of experimental design skills. *Chem. Educ. Res. Pract*., *24*(2)**,** 599–623. doi: 10.1039/D2RP00260D.

Szozda, A. R., Bruyere, K., Lee, H., Mahaffy, P. G. & Flynn, A. B. (2022) Investigating Educators' Perspectives toward Systems Thinking in Chemistry Education from International Contexts. *J. Chem. Educ*., *99*(7), 2474−2483. doi: 10.1021/acs.jchemed.2c00138.

Tosun, C. (2019). Scientific process skills test development within the topic ''Matter and its Nature'' and the predictive effect of different variables on 7th and 8th grade students' scientific process skill levels, *Chem. Educ. Res. Pract.*, ***20*(1)**, 160-174. doi: 10.1039/C8RP00071A.

Tseng, Y-J., Hong, Z-R., & Lin, H-s. (2022). Advancing students' scientific inquiry performance in chemistry through reading and evaluative reflection. *Chem. Educ. Res. Pract*., *23*(3), 616–627. doi: 10.1039/D1RP00246E.

Underwood, S. M., Posey, L. A., Herrington, D. G., Carmel, J. H., & Cooper, M. M. (2018). Adapting Assessment Tasks to Support Three-Dimensional Learning. *J. Chem. Educ*., *95*(2), 207–217. doi: 10.1021/acs.jchemed.7b00645.

van Brederode, M. E., Zoon, S. A., & Meeter, M. (2020). Examining the effect of lab instructions on students' critical thinking during a chemical inquiry practical. *Chem. Educ. Res. Pract*., *21*(4), 1173−1182. doi: 10.1039/D0RP00020E.

Vedder-Weiss, D., & Fortus, D. (2011). Adolescents' Declining Motivation to Learn Science: Inevitable or Not? *J. Res. Sci. Teach*., *48*(2), 199–216. doi: 10.1002/tea.20398.

Vedder-Weiss, D., & Fortus, D. (2013). School, teacher, peers, and parents' goals emphases and adolescents' motivation to learn science in and out of school. *J. Res. Sci. Teach*., *50*(8), 952–988. doi.org/10.1002/tea.21103.

Walker, J. P., & Sampson, V. (2013). Learning to Argue and Arguing to Learn: Argument-Driven Inquiry as a Way to Help Undergraduate Chemistry Students Learn How to Construct Arguments and Engage in Argumentation During a Laboratory Course. *J. Res. Sci. Teach*., *50*(5), 561−596. https://doi.org/10.1002/tea.21082.

Wang, Y., & Lewis, S. E. (2022). Towards a theoretically sound measure of chemistry students' motivation; investigating rank-sort survey methodology to reduce response style bias. *Chem. Educ. Res. Pract*., *23*(1), 240-256. doi: 10.1039/D1RP00206F.

Watts, F. M., & Finkenstaedt-Quinn, S. A. (2021). The current state of methods for establishing reliability in qualitative chemistry education research articles. *Chem. Educ. Res. Pract*., *22*(3), 565–578. doi: 10.1039/D1RP00007A.

Wheatley, K. (2018). Inquiry-Based Learning: Effects on Student Engagement. Honors Projects. 417., https://core.ac.uk/download/pdf/234759864.pdf Accessed: April 11. 2024.

Wren, D., & Barbera, J. (2013). Gathering Evidence for Validity during the Design, Development, and Qualitative Evaluation of Thermochemistry Concept Inventory Items. *J. Chem. Educ*., *90*(12), 1590–1601. doi: 10.1021/ed400384g.

Xu, H., & Talanquer, V. (2013), Effect of the Level of Inquiry of Lab Experiments on General Chemistry Students' Written Reflections. *J. Chem. Educ*. *90*(1), 21−28. doi: 10.1021/ed3002368

Zhang, J., & Zhou, Q. (2023). Chinese chemistry motivation questionnaire II: adaptation and validation of the science motivation questionnaire II in high school students. *Chem. Educ. Res. Pract*., ***24*(1),** 369–383. doi: 10.1039/D2RP00243D

Appendix

Table 1

Topics and context of the worksheets and teacher guides used in the school year 2022/2023

No	Topic	Experiments that Group 1 and Group 2 pupils had to do following step-by-step instructions, but Group 3 pupils had to design before doing the experiment	Context and elements of systems thinking in the "Let's think!" parts for motivation purposes. These are the same on the student worksheets of all the three groups.
7.	Reactivity series of metals and hydrogen, (redox reaction, electron transfer)	Pupils are given household hydrochloric acid, a piece of copper wire, a piece of aluminium foil, 2 test tubes or other containers and tweezers. Group 3 pupils have to design experiments to find out whether copper and aluminium are on the left or right side of hydrogen in the reactivity series. They are told that the metals and hydrogen are positioned in the reactivity series in the order of decreasing reducing power, from left to right.	Unlike the oxide layer on the surface of aluminium, the rust that forms on the surface of iron cannot protect it from the environment because it does not form a solid protective layer. Therefore, iron must be protected from rusting, for example by metal coatings made of tin or zinc. Pupils have to decide (using the reactivity series) whether such coatings can (theoretically) be produced by immersing the iron sheet in a solution containing tin (II) ions or zinc ions.
8.	Hardness of water, use of water softeners (precipita-tion reactions)	Pupils are given mineral water with high calcium ion content, trisodium phosphate, washing soda, soap solution, 3 test tubes with stoppers, 2 beakers, a measuring cylinder, 2 laboratory spoons, a Pasteur pipette and a ruler. Pupils watch a video to demonstrate the effect of cations that cause the hardness of water by measuring the height of the soap foam after shaking. Using a table that shows which anions form precipitates with the various cations, Group 3 pupils must work out which compounds could be used as water softener and show this by experiment.	1. According to several websites, commonly used household baking soda ($NaHCO_3$) is also suitable for water softening. After watching a video about the experiment, pupils have to decide whether this is true or false. 2. Pupils are explained that carbon dioxide dissolves in water and reacts with calcium carbonate in limestone. This converts it into water-soluble calcium bicarbonate, forming hard water. When the hard water loses some of its carbon dioxide content, the calcium carbonate precipitates as limescale or in form of stalactites. Pupils have to work out which of these processes is favoured by a rise or fall in temperature.
9.	Modelling industrial processes (production and use of quicklime and slaked lime)	Pupils are given distilled water, a piece of limestone, another piece of stone that is not limestone, phenolphthalein solution, 3 beakers or glasses, a Pasteur pipette, tweezers, an alcohol burner and matches. Pupils are explained how calcinated or quicklime /CaO/ and slaked or hydrated lime /Ca(OH)2/ are prepared of limestone. Pupils look at the reaction equations to understand that slaked lime is an alkaline substance. They then have to decide which of the stones on the trays is limestone that could be used to produce quicklime.	Pupils are explained that slaked lime mixed with sand and water forms mortar that can be used to fix bricks or plaster walls. In both cases, the slaked lime binds carbon dioxide in the air while converting into calcium carbonate and producing water. Next, pupils have to decide whether water and carbon dioxide are used or produced in the different steps of the process of making and using slaked lime.
10.	Modelling environ-mental processes (acid base reactions, pH, effects of acid rain)	Pupils are given tap water, pulverized limestone, sand, vinegar, red cabbage juice, 3 glasses, 2 Pasteur pipettes or eye droppers, 2 (lab) spoons. Pupils are explained that acid rain is mainly caused by the combustion of sulphur-containing carbon. They then need to investigate how the pH change in the lake water caused by acid rain is affected by the material (limestone or sandstone) on the lakebed.	Pupils are explained that calcium carbonate is the main component of limestone and the skeleton of calcareous aquatic organisms because limestone was formed from these organisms. Pupils are then asked to work out how acid rain affects the living conditions of calcareous animals (e.g. mussels, snails, corals) living in natural waters and how this effect may be influenced by the amount of sulphur-containing carbon burnt by humans.

11.	Modelling qualitative analysis (nutrients, health and diet)	Pupils are given granulated sugar, glucose, birch sugar (xylitol), 3 test tubes, test-tube rack, alcohol or Bunsen burner, matches, watch glass or ashtray. Pupils are explained that sugars can be caramelized, and the topping of a famous cake is made of that shiny caramel. Pupils are then shown that it is written on a website that a caramel cake topper for diabetics is made from birch sugar. Pupils have to decide whether this is true or false (i.e. whether birch sugar is really sugar or not).	Pupils are explained the dangers of diabetes when blood glucose levels are higher than 3.5-6 millimoles per litre. They also read that insulin lowers blood glucose levels. On the other hand, they learn that it is important to have enough glucose in the blood at all times. In case of stress, when cells use up a lot of glucose from the blood, blood glucose level drops. Glucose must then be replaced from the liver. Finally, using a few words and numbers given on the worksheet, pupils should complete a diagram showing how these opposing processes keep the blood glucose level within the range given.
12.	Plastics - pros and cons (household waste, raising environ-mental awareness)	Pupils are given 2 beakers/glasses with 100-100 cm^3 distilled water, 2 beakers/glasses with 0.1-0.1 g superabsorbent polymer (SAP), 1 measuring cylinder, 1 glass rod, 1 stand with clamp and ring, 1 glass funnel, 2 filter papers, 1 (lab) spoon, stopwatch/mobile phone with stopwatch function, 1 g sodium chloride. Pupils are told that superabsorbent polymers (abbreviated as "SAP") used in disposable paper nappies and sanitary pads are nowadays considered by many people indispensable. These plastics can absorb up to several hundred times their weight in various water-containing liquids. Then pupils have to determine whether the SAP in nappies and pads can absorb larger volumes of distilled water or body fluids such as urine or blood.	Pupils are told that in recent years more and more parents are choosing washable nappies to protect the environment. However, many people question whether these products are really environmentally friendly, for example because of the potential waste of water during washing. Pupils are given data to calculate whether using washable nappies requires more or less water than disposable nappies, generates more or less waste and is therefore a more or less environmentally friendly choice. They should also calculate which is more cost-effective. Finally, they have to decide which option they would choose.

Table 15

The Cohen's d Effect Size Values Calculated from the Means and Standard Deviations of the Differences Between the Test Scores (T1 - T0; T2-T1 and T2-T0) for the Whole Test ("TOTAL"), the DCK Tasks ("") and the EDS Tasks ("EDS") (N=756)

	TOTAL			DCK			EDS		
	T1 – T0	T2 – T1	T2 – T0	T1 – T0	T2 – T1	T2 – T0	T1 – T0	T2 – T1	T2 – T0
Group 2/ Group 1	-0.19	0.29	0.09	-0.25	0.34	0.09	-0.05	0.12	0.05
Group 3 / Group 1	0.32	-0.03	0.26	0.06	0.11	0.14	0.40	-0.16	0.26
Group 3 / Group 2	0.51	-0.25	0.19	0.28	-0.20	0.06	0.44	-0.31	0.21

Table 16

The Effects of the Assumed Parameter "Group" (Instruction Methods) Estimated by the Model of the ANCOVA Analysis (Absolute Mean Scores) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK"), the EDS Tasks ("EDS") and the Significance of their Differences for the Three Tests (N=756)

Group	$T0_{TOTAL}$	$T1_{TOTAL}$	$T2_{TOTAL}$	$T0_{DCK}$	$T1_{DCK}$	$T2_{DCK}$	$T0_{EDS}$	$T1_{EDS}$	$T2_{EDS}$
Group 1	11.01	8.90	9.00	5.32	4.40	3.43	5.69	4.47	5.55
Group 2	11.06	8.26	9.40	5.37	3.91	3.66	5.69	4.32	5.73
Group 3	10.55	9.75	9.0	5.25	4.41	3.71	5.311	5.24	6.02
Significant difference*	-	2 < 1 < 3	1 < 3	-	2 < 1, 3	-	-	1, 2 < 3	1 < 3

Note: * $p<0.05$

Table 17

The Effects of the Assumed Parameter "School Ranking" Estimated by the Model of the ANCOVA Analysis (Absolute Mean Scores) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK"), the EDS Tasks ("EDS") and the Significance of their Differences for the Three Tests (N=756)

School ranking	$T0_{TOTAL}$	$T1_{TOTAL}$	$T2_{TOTAL}$	$T0_{DCK}$	$T1_{DCK}$	$T2_{DCK}$	$T0_{EDS}$	$T1_{EDS}$	$T2_{EDS}$
1. Low	9.76	8.30	8.98	5.13	3.64	3.69	4.63	4.47	5.11
2. Medium	10.43	8.70	9.50	5.06	4.15	3.51	5.37	4.44	5.91
3. High	12.43	9.91	9.73	5.74	4.94	3.61	6.69	5.12	6.27
Significant difference*	1 < 2 < 3	1, 2 < 3	1 < 3	1, 2 < 3	1 < 2 < 3	-	1 < 2 < 3	1, 2 < 3	1 < 2 < 3

Note: * $p<0.05$

Table 18

The Effects of the Assumed Parameters (Sources) and the Covariate ("Prior Knowledge", T0) on the Changes in Test Scores (T1 - T0, T2 - T1, T2 - T0) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK") and the EDS Tasks ("EDS") in the Beginning of the Project (T0), in the End of Grade 7 (T1) and in the End of Grade 8 (T2) (N=756)

Parameter (Source)	PES (Partial Eta Squared)								
	TOTAL			DCK			EDS		
	T1 – T0	T2 – T1	T2 – T0	T1 – T0	T2 – T1	T2 – T0	T1 – T0	T2 – T1	T2 – T0
Group	0.042*	0.023*	0.012*	0.018*	0.018*	0.004	0.040*	0.014*	0.014*
School ranking	0.046*	0.015*	0.009	0.079*	0.057*	0.001	0.023*	0.019*	0.059*
Mother's education	0.004	0.001	0.001	0.002	0.006	0.002	0.005*	0.000	0.010*
Gender	0.000	0.002	0.000	0.000	0.000	0.000	0.001	0.003	0.002
Prior knowledge (T0)	0.333*	0.003	0.365;	0.335*	0.000	0.270*	0.464*	0.005	0.565*

Note: * Significant at $p < 0.025$ level (Bonferroni correction

Table 19

The Values Estimated by the ANCOVA Model Showing the Effect of the Assumed Parameter "Group" (Instruction Methods) on Changes in Performance (T1 – T0, T2 – T1, T2 – T0) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK") and the EDS Tasks ("EDS") (N=756)

Group	$T1_{TOTAL}$ – $T0_{TOTAL}$	$T2_{TOTAL}$ – $T1_{TOTAL}$	$T2_{TOTAL}$ – $T0_{TOTAL}$	$T1_{DCK}$ – $T0_{DCK}$	$T2_{DCK}$ – $T1_{DCK}$	$T2_{DCK}$ – $T0_{DCK}$	$T1_{EDS}$ – $T0_{EDS}$	$T2_{EDS}$ – $T1_{EDS}$	$T2_{EDS}$ – $T0_{EDS}$
Group 1	-2.40	0.09	-2.31	-1.19	-0.97	-2.16	-1.25	1.08	-0.17
Group 2	-3.05	1.15	-1.90	-1.67	-0.25	-1.92	-1.40	1.41	0.01
Group 3	-1.56	0.05	-1.51	-1.18	-0.70	-1.88	-0.48	0.77	0.30
Significant difference*	2 < 1 < 3	1, 3 < 2	1 < 3	2 < 1, 3	1, 3 < 2	-	1, 2 < 3	3 < 2	1 < 3

Note: * Significant at $p < 0.025$ level (Bonferroni correction)

Table 20

The Values Estimated by the ANCOVA Model Showing the Effect of the Assumed Parameter "School Ranking" on Changes in Performance (T1 – T0, T2 – T1, T2 – T0) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK") and the EDS Tasks ("EDS") (N=756)

School ranking	$T1_{TOTAL}$ – $T0_{TOTAL}$	$T2_{TOTAL}$ – $T1_{TOTAL}$	$T2_{TOTAL}$ – $T0_{TOTAL}$	$T1_{DCK}$ – $T0_{DCK}$	$T2_{DCK}$ – $T1_{DCK}$	$T2_{DCK}$ – $T0_{DCK}$	$T1_{EDS}$ – $T0_{EDS}$	$T2_{EDS}$ – $T1_{EDS}$	$T2_{EDS}$ – $T0_{EDS}$
1. Low	-3.01	0.68	-2.33	-1.95	0.05	-1.90	-1.25	0.64	-0.61
2. Medium	-2.61	0.80	-1.81	-1.43	-0.64	-2.08	-1.28	1.47	0.19
3. High	-1.40	-0.18	-1.58	-0.65	-1.33	-1.98	-0.48	1.15	0.55
Significant difference*	1, 2 < 3	3 < 1, 2	1 < 3	1 < 2 < 3	3 < 2 < 1	-	1, 2 < 3	1 < 2, 3	1 < 2 < 3

Note: * Significant at p < 0.025 level (Bonferroni correction)

Table 21

The Values Estimated by the ANCOVA Model Showing the Effect of the Assumed Parameter "Mother's Education" on Changes in Performance (T1 – T0, T2 – T1, T2 – T0) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK") and the EDS Tasks ("EDS") (N=756)

Mothers' education	$T1_{TOTAL}$ – $T0_{TOTAL}$	$T2_{TOTAL}$ – $T1_{TOTAL}$	$T2_{TOTAL}$ – $T0_{TOTAL}$	$T1_{DCK}$ – $T0_{DCK}$	$T2_{DCK}$ – $T1_{DCK}$	$T2_{DCK}$ – $T0_{DCK}$	$T1_{EDS}$ – $T0_{EDS}$	$T2_{EDS}$ – $T1_{EDS}$	$T2_{EDS}$ – $T0_{EDS}$
1. No degree in higher education	-2.61	0.60	-2.01	-1.47	-0.40	-1.87	-1.24	1.04	-0.20
2. Has a degree in higher education	-2.07	0.26	-1.81	-1.22	-0.88	-2.10	-0.84	1.14	0.29
Significant difference*	-	-	-	-	2 < 1	-	-	-	1 < 2

Note: * Significant at p < 0.025 level (Bonferroni correction)

Table 22

The Values Estimated by the ANCOVA Model Showing the Effect of the Assumed Parameter "Gender" on Changes in Performance (T1 – T0, T2 – T1, T2 – T0) for the Whole Test ("TOTAL"), the DCK Tasks ("DCK") and the EDS Tasks ("EDS") (N=756)

Gender	$T1_{TOTAL}$ – $T0_{TOTAL}$	$T2_{TOTAL}$ – $T1_{TOTAL}$	$T2_{TOTAL}$ – $T0_{TOTAL}$	$T1_{DCK}$ – $T0_{DCK}$	$T2_{DCK}$ – $T1_{DCK}$	$T2_{DCK}$ – $T0_{DCK}$	$T1_{EDS}$ – $T0_{EDS}$	$T2_{EDS}$ – $T1_{EDS}$	$T2_{EDS}$ – $T0_{EDS}$
1. Boy	-2.29	0.35	-1.95	-1.32	-0.61	-1.93	-0.99	0.97	--0.03
2. Girl	-2.38	0.52	-1.87	-1.37	-0.66	-2.04	-1.09	1.21	0.12
Significant difference*	-	-	-	-	-	-	-	-	-

Note: * Significant at p < 0.025 level (Bonferroni correction)