SIEMENS Stiftung

Evaluation guidelines for the education program Experimento

siemens-stiftung.org/Experimento

Published by:
Siemens Stiftung
Kaiserstrasse 16
80801 Munich
Tel.: +49 (0) 89 54 04 87-0
Fax: +49 (0) 89 54 04 87-440
info@siemens-stiftung.org
www.siemens-stiftung.org

Authors
Prof. Dr. Eva-Maria Lankes, TUM School of Education, Munich
Dr. Heidi Haslbeck, TUM School of Education, Munich

In collaboration with
Prof. Dr. Birgit H. Neuhaus, LMU Munich
Dr. Lucia Kohlhauf, LMU Munich
Prof. Dr. Lena von Kotzbue, Universität Salzburg
Dr. Barbara Filtzinger, Siemens Stiftung Munich

1st edition

As a nonprofit foundation, we promote sustainable social development, which is crucially dependent on access to basic services, high-quality education, and an understanding of culture. To this effect, our project work supports people in taking the initiative to responsibly address current challenges. Together with partners, we develop and implement solutions and programs to support this effort, with technological and social innovation playing a central role. Our actions are impact-oriented and conducted in a transparent manner.

Foreword

These evaluation guidelines provide Siemens Stiftung partners in the target regions of Europe, Latin America, and Africa with evidence-based criteria for evaluating their Experimento programs. It is meant to inspire systematic observation of work done on the ground, leading to questions, reflection, change, and growth.

From 2018-2020, an evaluation of the education program Experimento was conducted in Germany using the tools in this handbook. Many of the evaluation instruments were developed specifically for that purpose. This led to the creation of similar instruments and tools for evaluating localized Experimento adaptation in Argentina, Brazil, Chile, Colombia, Kenya, Mexico, Nigeria, Peru, and South Africa.

The evaluation handbook is the result of a collective effort by several people. We would like to take this opportunity to thank those who joined us in this undertaking with their assistance, guidance, inspiration, and support.

We also extend our thanks to colleagues at the Institute for Biology Education at Ludwig-Maximilians-Universität in Munich – Prof. Dr. Birgit E. Neuhaus, Prof. Dr. Lena von Kotzebue (now at Universität Salzburg) and Dr. Luzia Kohlhauf – for their close collaboration and professional support during all phases of the project. All three provided expertise and didactic insights into developing the checklist for the assessment of materials, tests on educators' didactic knowledge, tests on experiments for schoolchildren, and coding guidance for video assessments of lessons.

Our thanks go as well to the employees, research assistants, and students at Technische Universität München who worked on the project as part of their duties at the university or in pursuit of their dissertations. All illustrations in the test exercises were created by Stefanie Pielmeier.
As part of her master's thesis, Pauline Böttcher-Graf worked on the development and pilot phase of the educator survey on self-concept in science subjects. Franziska Tezcan and Laura Müller help develop the coding guidance for the video analysis. Sina Hafner's bachelor thesis involved preparatory work on creating survey items for the variable control strategy.

Special thanks to colleagues from other research institutes for providing their tried and tested scientific instruments in the pilot phase and now for this set of guidelines:

- Prof. Dr. Andreas Vorholzer, Institute of Physics Education at Justus Liebig Universität Gießen (Test exercises on educators' methodological knowledge)

- Dr. Julia Arnold, research associate at Center for Science and Technology Education at Fachhochschule Nordwestschweiz (Test exercises on educators' methodological knowledge)

- Prof. Dr. Kornelia Möller, Institute for Didactics of Science Education at Westfälischen Wilhelms Universität Münster (Coding guidance for classroom observation)

- Dr. Andrea Saffran, research associate at Chair for Developmental Psychology at Ludwig-Maximilians-Universität Munich (Test exercises on student experimentation competence)

- Junior professor Dr. Martin Schwichow, Physics education at Pädagogische Hochschule Freiburg (Test exercises on student experimentation competence)

- Prof. Dr. Alexander Kauertz (Universität Koblenz-Landau), Prof. Dr. Thilo Kleickmann (Christian-Albrecht-Universität Kiel), Prof. Dr. Kim Lange-Schubert (Universität Leipzig), Dr. Annika Ohle-Peters (Technische Universität Dortmund), Dr. Steffen Tröbst (IPN Kiel), Prof. Dr. Hans Fischer (Universität Duisburg-Essen), Prof. Dr. Kornelia Möller, Dr. Anne Ewerhardy, Dr. Katharina Fricke, Dr. Katharina Pollmeier, Dr. Lena Walper (all Westfälische Wilhelms-Universität Münster) (Self-concept and scientific interest survey)

To Angela Clerc and Christa Mühlbauer from Siemens Stiftung, we say thanks for your patience and generous support. This project could not have been completed without it.

We hope that this evaluation handbook contributes to further developments in science education within the international education program Experimento. We would like to wish the best of luck to anyone putting these guidelines to use.

Prof. Dr. Eva-Maria Lankes	Dr. Heidi Haslbeck	Dr. Barbara Filtzinger
Chair of School Pedagogy	Project Management	Head of Unit Education,
at Technische Universität Munich	Experimento Evaluation	Siemens Stiftung

Evaluation guidelines for the education program Experimento

Contents

1 Evaluation guidelines for the education program Experimento

“As not everything can be done, there must be a basis for deciding which things are worth doing. Enter evaluation.” (Patton 2008, p. 16).

Actions require resources: time, money, and effort. Anyone hoping to use resources effectively will want to reassure themselves that the proper actions are being put to good use. These guidelines are meant to help determine if Experimento has achieved its goals in a specific context while also providing indicators of what can be improved.

1.1 For whom have these guidelines been created?

Siemens Stiftung provides a wide range of materials for conducting experiments in science lessons. These include teacher handouts, worksheets for students, and materials needed to conduct experiments. The experimentation materials are often provided in a box, with individual materials for every person in a class, in some cases. In partner countries and regions, the materials are adapted to fit various geographic or educational circumstances. Teachers learn how to use the materials in their lessons in locally-administered teacher training seminars.

Siemens Stiftung commissioned Technische Universität Munich, in collaboration with Ludwig-Maximilians-Universität Munich, to come up with instruments for evaluating Experimento and to outline their use in these guidelines.

The guidelines provide local Siemens Stiftung partners (ministries, national institutes, universities, etc.) with evidence-based criteria for evaluating Experimento in schools. They are meant to inspire systematic observation of work done on the ground, leading to questions, reflection, change, and growth. The guidelines describe quality standards with common evaluation criteria for the use and deployment of Experimento in each region. These include examples of evaluated instruments, insights into their use, and methods for analyzing and interpreting results.

1.2 What will you find in these guidelines?

- Are you interested in reflecting on the goals of Experimento and gaining confidence that you are on the right track? Then read chapter two – The Experimento education program.
- Are you looking for information about planning, implementing, and assessing an evaluation? Then read chapter three – Evaluation as a quality assurance method.
- Have you adapted Experimento handouts and worksheets for your country and want to know if the adapted materials meet the education program’s quality? Head over to chapter five – How Useful are the Handouts, Student Worksheets, and Experimentation Materials?
- Do you offer regular seminars for teachers on the use of Experimento and want to know if these seminars are serving their purpose? Chapter six is for you – Quality of Teacher Training?
- Have you disseminated materials to educators and now want to know how often and in what ways these materials are utilized? Then read chapter seven – How do Teachers use Experimento?
- Do you want to gain a sense of how Experimento is used in real learning situations? You should explore chapter eight – Quality Lessons with Experimento.

- Would you like to show teachers how they can check and evaluate student competence for experimentation? Then read chapter nine – What do Students Learn from Scientific Experimentation?
- Are you interested in fully understanding an evaluation of the education program Experimento? Then read everything.

2 The education program Experimento

The international educational program Experimento supports science learning around the world, from early childhood all the way to the end of a child's time at school.

2.1 Why is it important to learn about science at an early age?

Science, technology, engineering, and mathematics (STEM) shape and change our world. We deal with these issues every day in our personal and professional lives. Knowledge of technological and scientific concepts and contexts, in addition to the ability to use this knowledge responsibly, are important determining factors for how we conduct our lives and for the development and well-being of a society (Hazelkorn et al., 2015).

More than ever before, the future of global development is dependent on young people finding their place in a world shaped by technology and science. These young people must be prepared to face this reality – now and in the future – responsibly and constructively. However, the specific skills needed for young people to embrace this role have been lacking in many areas for years. (Milberg, 2009). Interest in science ebbs over the course of a scholastic education (Krapp & Prenzel, 2011).

At the same time, studies show that exposure to science in early childhood education influences an individual's educational biography and later school performance (Bybee & Fuchs, 2006). This is the reason early childhood science education is encouraged and supported by many initiatives. Students should acquire real-life knowledge about science and technology as early as possible in preparation for occupational profiles that do not yet exist.

2.2 Why is experimentation important for quality science education?

The subject matter that is taught in science lessons already exists; it could be learned from a book. So why is experimentation relevant for quality science lessons?

One answer stands out among the many that exist: because it's fun. Experiments are vivid and serve as a first step for many activities. Experimentation creates joy from the scientific method and generates interest in scientific issues, fulfilling an important goal of science education: establishing an interest in science that is long-lasting and extends beyond the classroom.

But experimentation does much more: When done properly, it opens children's eyes to their surroundings. They begin to understand why things are the way they are by developing analytic and rational perspectives. They learn to distinguish between "believing" and "knowing" and develop the ability to think critically. Finally, they grasp the idea that actions may have consequences, and that these should be considered and discussed. These lessons are essential for basic science education and are fundamental principles for conscientious engagement with our environment (Abd-El-Khalick et al., 2004; Blanchard et al., 2010).

2.3 What are the goals of the Experimento program?

The pursuit of early childhood science education is the focus of Siemens Stiftung's international education program Experimento, which utilizes real-life examples of science and technology learning content for classroom lessons.

Experimento supports teachers and educators in implementing experiments in inquiry-based and problem-based learning situations. In seminars, handouts, and other materials, the program seeks to generate interest among teachers for science education, share knowledge and didactic competencies, and instill confidence in conducting experiments.

Children from preschool upwards can use Experimento to learn about energy, the environment, and health at their own pace. In addition to acquiring knowledge about nature and technology, they also learn scientific methods and develop social skills, values, attitudes, and language abilities while building personal interests and confidence.

With Experimento, Siemens Stiftung pursues impact-oriented science and technology education that encourages individual learning potential, from preschool until graduation. Social and conscientious behavior and attitudes are developed by connecting content to value-based issues.

2.4 What does Experimento offer?

Experimento was developed by education experts for use in preschools, primary schools, and secondary schools. It targets teachers and educators at all three levels of school education.

Experimento provides teachers with an age-appropriate selection of materials on energy, environment, and health that fit curricular criteria. It also provides guidelines for how to teach using experiments and how the experiments can be integrated into classroom lessons.

The Siemens Stiftung Media Portal contains around 130 experiments developed by experts for Experimento for age groups 4 – 7 (Experimento | 4+), 8 – 12 (Experimento | 8+) and 10 – 18 (Experimento | 10+). The materials are available in German, English, and Spanish and can be downloaded free of charge from the website under an open CC BY-SA 4.0 license. This indicates that if the original author is named, all teaching materials can be used, modified, combined, and republished under the same licensing conditions (https://medienportal.siemens-stiftung.org/de/home).

The portal contains a folder with teacher handouts for educators in each age group. The handouts contain basic pedagogic concepts, general safety and hygiene measures, and a list of all materials needed for an experiment.

Each of the three subjects (energy, environment, and health) also contains an introduction and subject-specific safety instructions. Finally, the handouts contain the relevant subject matter expertise in each field, descriptions of all experiments, and didactic tips for implementing the experiments in learning situations. The handouts for Experimento | 8+ and 10+ also include student worksheets.

Participants in the training seminars receive the materials needed to conduct each of the experiments. The materials should be stored for easy access, ensuring they can be used by as many teachers at a school as possible.

Since 2021, Experimento is being offered in a “blended learning” format. This allows the world of Experimento to be explored using alternating online and in-person sequences. The following online modules are available in German: https://medienportal.siemens-stiftung.org/de/fortbildungen.

2.5 What didactic principles are behind Experimento?

Enthusiasm and curiosity spark children’s interest in nature and technology at an early age and maintain it across all age groups. Experimento is based on the principle of inquiry-based learning: Through experiments, children gain the opportunity to discover scientific and technological interrelationships for themselves, which become permanently engrained. Teachers and educational experts achieve this by guiding children as they construct their own view of the world through active participation in the learning process (Labudde & Möller, 2012; Müller et al., 2016).

Teachers and educational experts guide children as they construct their own view of the world:

- Experiments are related to age-appropriate phenomena and experiences drawn from everyday life (Anders et al., 2013).
- Experiments solve a specific problem or address a specific question.
- Experiments help children to challenge their ideas and prior knowledge and, potentially, change or expand them.
- Experiments provide opportunities for understanding different viewpoints and perspectives and exposure to certain values.

Teachers and educational experts are actively involved in the children’s learning process:

- They talk with children about their ideas, perceptions, and prior knowledge regarding a phenomenon or a problem.
- They let the children work on their own as much as possible, allowing them to examine their perceptions, try things out, and experiment.
- They give children the opportunity to talk about their experiences with each other, to learn from each other, and to reflect on their learning process.

Experimento supports teachers and educational experts in practically and effectively embedding experiments into learning situations. The handouts for teachers and the worksheets for students provide materials that balance instruction and construction. It shows educators how students can become autonomous learners by conducting experiments.

3 Evaluation as a quality assurance method

“Evaluation is the systematic examination of use and/or value of a subject (of evaluation) based on empirical data.” (DeGEval, 2016, p. 66). “Evaluation in a pedagogical context” covers “pedagogic actions or processes, institutions, or systems in the education system.” (Brunner, Stanat & Pant, 2014, p. 502).
Typical subjects of evaluations are specific interventions or broad programs that pursue certain individual and collective changes (Döring & Bortz 2016, p. 980). A summative evaluation examines the effectiveness of an intervention upon its conclusion and informs decision-making.

For the evaluation of Experimento, our concept utilizes a formative evaluation that accompanies processes for optimizing in-progress interventions.

3.1 What is the purpose of an evaluation in an education context?

“An evaluation conducted in an education context should serve to assess the quality of pedagogic interventions, education institutions, and entire education systems. Ideally, the evaluation should initiate interventions or optimization efforts of what has been evaluated” (Lüftenegger, Schober & Spiel, 2019, p. 518). Quality is measured against previously defined expectations or goals that are meant to be achieved through a particular intervention. These are discussed among the relevant stakeholders prior to the evaluation. “This is the only way it is possible to gain comprehensive understanding of the evaluation subject and to ensure that the results of the evaluation are also put to use later on” (Lüftenegger, Schober & Spiel, 2019, p. 520).

In the context of education interventions, the stakeholders are the teachers and educators carrying out and implementing them. But stakeholders also include students, and in a broader context, their parents, who are all affected by specific interventions. Stakeholders also include those who are responsible for creating, implementing, and optimizing these interventions, such as developers, contracting entities, accountable persons, and decision makers (Döring & Bortz, 2016). Within a formative evaluation, stakeholder groups can engage in a dialog and work on developing interventions collaboratively.

3.2 What is the subject of evaluation in a pedagogical context?

Over the course of a pedagogical intervention, there are a variety of factors that can influence the result and/or the intervention’s impact. Stufflebeam & Shinkfield (2007) differentiate between context, input, process, and product (the CIPP model). Each of these aspects can be subjects of evaluation. Context describes the environment in which an intervention is to be implemented. Input refers to the resources (material and personnel) that are necessary for implementing the intervention. How the intervention is carried out is known as the process, and product describes the result.

In our evaluation criteria, we propose examining multiple aspects pertaining to the quality of the education program Experimento:

- Our **input to the pedagogical intervention** is the provision of instruments for assessing the quality of teacher handouts and experimentation materials (see chapter 5) in addition to the quality of teacher training seminars (see chapter 6).
- The **quality of the process** is analyzed through documentation in a logbook (see chapter 7) and with classroom observation (see chapter 8).
- We consider students’ abilities and competencies in experimentation to be the **product of the intervention** (see chapter 9).

An evaluation of *context,* or the framework of an intervention, occurs before the intervention is planned. Framework conditions in the education systems of Experimento focus regions vary significantly. Influencing these conditions is outside the scope of the Experimento program, which is why we decided against an evaluation of framework conditions (context) for the purposes of these guidelines.

Depending on what is being asked, an evaluation can be a time and resource-consuming process for those surveyed. In addition, some evaluation questions require a degree of expertise for a thorough answer. Therefore, an evaluation should always be carefully targeted and implemented sparingly. A simple acceptance survey can be conducted using original questions with little knowledge of survey methods. For more complex survey queries – such as the quality of materials, the impact of a teaching seminar, or the constellation of a classroom situation – it is better to find experienced partners and utilize tried and tested instruments, like those described in these evaluation guidelines.

Simply considering the timing and frequency of a survey can help save resources:

- The acceptance of an intervention, whether it is teacher seminars or the usefulness of materials, is quickly established. This can serve as an early indicator of whether an intervention is generally liked or not.
- The assessment of the quality of a handout is time-consuming and should only take place when an update of the handout is already planned – the quality analysis gives indications about the parts of the handout that should be updated.
- For individual educators, the beginning of the school year is the best time to gain a sense of where students are, what they are capable of, and where additional learning opportunities might make sense. After establishing the status quo, repeating an evaluation after certain lessons can help identify progress or change.
- The assessment of real-world use of the experiment materials in lessons, including the frequency of each experiment, is an important question when it comes to the allocation of resources, especially when determining the right time to acquire new experimentation materials.
- An analysis of teaching practices through observation is always a resource-intensive exercise that requires a high degree of expertise to execute. However, such an analysis enables an in-depth and detailed evaluation of teacher performance in the classroom, which provides valuable insights into how teacher training and handouts should be structured.

3.3 How should an evaluation be conducted?

An evaluation produces a result that indicates whether an intervention's goals have been achieved.

To arrive at this result, the following steps are necessary:
- Defining goals to be achieved by the intervention
- Operationalizing goals and identifying indicators
- Making decisions about data collection and acquisition methods
- Conducting data assessment and interpretation

3.3.1 Defining goals

Implementing an intervention, such as the intention to conduct experiments with students in the classroom, is done with a certain result or impact in mind. In the pedagogical context, this is usually about multidimensional goals containing cognitive, social, and motivational aspects.

The goals of Experimento pedagogical interventions are defined by the broader goals of the program itself (see chapter 2.3):
- In addition to learning about nature and technology, students should acquire methodical abilities and develop social skills, values, attitudes, and language abilities while building their interests and self-confidence. Within the CIPP model, these outcomes would be the intervention's *product.*
- Using the materials in classroom lessons (process), in the manner defined by Experimento, is designed to develop the competencies and attitudes in students that are listed above: Teachers use experiments in inquiry- and problem-based learning situations in a way that gives students the opportunity to discover science and technology interrelationships on their own and to permanently internalize this knowledge. Children form their own views of the world and become an active part of the learning process.
- With seminars, handouts, and other materials (input), the program supports teachers in shaping the (learning) process: It teaches knowledge and didactic competencies, kindles interests, and increases teachers' confidence in their ability to conduct experiments.

3.3.2 Operationalizing goals and identifying indicators

Before beginning an evaluation, a more precise determination needs to be made about how goals will be defined as fulfilled or not. This involves breaking down general, overarching goals into smaller parts with specific and measurable indicators (parameters). These can then be used to measure progress on both specific and overarching goals. This process is called operationalization. The indicators should be as specific as possible, since they inform the creation of the instruments that will be used for measuring indicators. Success criteria can also be established for assessing results at the end of the evaluation.

Examples:

Goal **(product):**	Students should have more fun while learning.
Possible indicators:	Self-assessment of the joy of learning; increased student participation during lessons; ...
Success criteria:	Before/after comparison

Goal **(process):**	Teachers provide students with the opportunity to steer their own learning.
Possible indicators:	The use of answer keys for checking students' own work; ability during lessons to select from choices or make own decisions; freely-definable working stages; ...
Success criteria:	Increased use of the intervention compared to starting point. Stricter criteria can be used to define success more specifically: doubling the use of the intervention compared to starting point, for example.

Establishing indicators is the most challenging step in defining goals. It often involves a process of bouncing back and forth between (theoretically-founded) expectations and what is observable or measurable. Working in a team has proven to be useful in this process. Indicators that have been developed in theory need to be tested in practice. Sometimes it will be necessary to rethink certain goals. In some cases, it may be necessary to accept the fact that while every goal can be operationalized, not every goal can be measured.

3.3.3 Decisions about data collection and acquisition methods

An evaluation may make use of various methods of empirical social research (for an overview, see examples such as Cohen, Manion & Morrison, 2018).

One method of evaluation that is frequently used is the questionnaire survey. This is an easy way to take large samples. In addition, a survey can easily be adjusted to fit target or age groups (by changing the scope of the questions, using smileys, etc.) There are many providers available online that offer tools for creating a survey; these often include automated procedures for survey analysis. However, in a public sector setting, use of these tools should involve a close examination of their data protection guidelines (see also 3.5).

It is important to acknowledge that respondents to a survey will always be providing their personal views. This is, of course, relevant for assessing acceptance or motivation. However, social desirability distortions cannot always be filtered out, even when the survey is anonymous. A meaningful instrument for evaluating criteria such as classroom lessons is observation, whether this is done by a trained participatory observer or through a video recording of the lesson. Participatory observation and video analysis are resource-intense methods that should be used sparingly; they should be targeted and make use of questioning that is as straightforward as possible.

Knowledge tests are a good method for assessing learning outcomes. This is not only about completing exercises (the repetition of facts vs. demonstrating comprehension with examples and practical application). Knowledge tests can be administered online, but open-ended questions mean increased effort in coding answers. Multiple choice questions can be automatically assessed. For goals that pertain to the capacity to act rather than a pure assessment of knowledge, a combination of knowledge tests and observation is a good assessment tool.

A quantitative content analysis (such as the method Mayring, 2015) is a method to make qualitative data (obtained in written answers or interviews, for example) quantifiable and to evaluate it systematically. Using a standardized categorization system, statements in written text can be coded and collected.

A logbook is a survey method that looks at longer time intervals. It provides for input from specific contextual situations. Teachers are asked to fill out a short survey in a logbook each time they use experimentation materials in lessons.

For all instruments, it is recommended to begin with open-ended, qualitative methods, (such as interviews). Such methods establish appropriate evaluation criteria, relevant constructs, and participant-related items. Once the instruments have been developed, a pilot phase statistically assesses if the instruments measure what they are intended to measure. The instruments presented in this set of guidelines were all developed and piloted in this way.

3.3.4 Data assessment and interpretation

Every interpretation should begin with a descriptive assessment of frequency. For each parameter, this indicates the distribution of the answers, with mean values and standard deviations completing the picture. A description of the current state always precedes evaluation and interpretation: What does the data tell us? What do we see? What is the situation?

Only then can we ask additional questions: What does the data mean? Are we satisfied with it? Was the intervention successful? If an evaluation study was consequently and logically planned and executed, the answers to these questions are relatively easy to find: with appropriate indicators, the results will provide insights into the subgoals / partial constructs that can be applied to the overarching goal.

3.4 What are the standards of a good evaluation?

Four basic characteristics are used to determine the quality of an evaluation. Each characteristic can be described by a list of standards (DeGEval 2016, p. 18ff).

- **Utility:** “The utility standards are meant to ensure that the evaluation is used for agreed upon and declared purposes that, to the extent possible, is in the interest of supplying information to the intended users” (ibid., p 34). This question must be answered at the beginning of every evaluation – what should happen because of the results? Who receives them? Who should continue working on them? The effort involved in an evaluation can only be justified when the results lead to improvements. Utility should be as clear as possible for everyone involved.
- **Feasibility:** “Feasibility standards are meant to ensure that planning and execution of the evaluation is realistic, considered, diplomatic, and cost-conscious” (ibid., p. 39). This raises the question of what data is truly necessary to answer the questions, which depends on the methods of data collection. The effort involved must be in proportion to the available resources. Any pushback or criticisms must be considered and diplomatically handled to eliminate risks to the execution of the survey.
- **Fairness** (propriety): “Fairness standards are meant to ensure that all participants and impacted individuals and groups are treated fairly and with respect throughout the course of the evaluation” (ibid., p. 41). An evaluation serves the purpose of progress, which is why it needs the support and acceptance of all people and groups involved in the process. For these stakeholders, the willingness to participate is increased through respectful and fair

interpersonal interactions, transparency regarding the process and results, and inclusion when it comes to potential consequences or changes. In the education sector, it is particularly important to have a clear distinction between evaluation and judgement. Evaluation serves to improve an intervention, not to judge teachers or other stakeholders.

- **Accuracy:** “Accuracy standards are meant to ensure that an evaluation yields valid and comprehensible results on each subject of evaluation and line of questioning” (ibid., p. 44). This standard is about the appropriateness and quality of the methods used. Results will only be accepted if they are conducted by the book, based on objective and valid procedures. The people conducting the evaluation must account for the procedures used and for utilizing relevant professional methods.

3.5 Notes on data protection

Data protection refers to the security of personal information. If personal information is collected as part of a study, applicable data protection regulations must be upheld. Data protection regulations contain provisions pertaining to the voluntary nature of participation, anonymization of personal data, saving and deleting data, and publishing results.

When introducing a survey, participants should be informed of the applicable data protection regulations regarding:

- Voluntary participation
- Confidentiality of responses
- Processing data only if it has been made anonymous or if pseudonyms are used
- The planned date for data to be deleted
- The conditions for possible publication of the results

If data will be collected in multiple stages and later collated, it is recommended to implement a simple personal code, such as:

How to create your own personal code:
1st and 2nd letters of your mother’s first name,
1st and 2nd letters of your father’s first name,
Date of your birth.
Example:
A person’s mother is called Sonja, the father is Helmut, and their birthday is 8 June 1989.
The code would be: SOHE08

4 Overview of the Experimento evaluation modules

Experimento evaluations are primarily used to check if the program's regional offerings are appropriate for achieving the program's goals (see Chapter 2) and to encourage continued development of these offerings.

The evaluation criteria contained in this document introduces processes and instruments that are designed to help when reflecting on the various core principles of the program in a regional setting and, if necessary, to develop them further. Even if many of these aspects are applicable to an evaluation of the use of Experimento in preschools, the methods and instruments presented here are primarily designed for teachers utilizing Experimento in primary and secondary school lessons.

The program's basic components are:

- Experimento handouts
- Experimento teacher training seminars
- Experimento experimentation materials

The handouts and training seminars are meant to motivate educators to implement Experimento in the classroom and enable them to use the experiments to advance students' learning development in an effective and didactically-relevant way.

A checklist can assess the <u>quality of the handouts</u> (evaluation section 1, chapter 5). The quality of the handouts is measured against three criteria:

- Are the included experiments relevant and exemplary of fundamental science education?
- Does the material meet scientific, methodic, and didactic standards in the subject?
- Do the handouts make it easier for teachers to use experiments in the classroom (usability)?

The <u>impact of the Experimento training seminars</u> on teachers' mindsets and expertise can be assessed with a survey and a test (evaluation section 2, chapter 6):

- The survey provides insights into the convictions, interests, and self-concept that teachers bring to the seminars, and indicates whether their confidence grows over the course of the seminar.
  After the seminar, the survey contains questions about how satisfied participants were with
  - The seminar content and
  - The instructors and environment.
- The test can assess participants' prior knowledge of methods and didactics and determine if these change over the course of the seminar.

The <u>use of the materials</u> is documented by educators in a logbook (evaluation section 3, chapter 7). Assessing the logbooks provides insights into:

- How often the materials are used,
- Which experiments are conducted the most and least frequently,
- How often student worksheets are used,
- Grade levels in which the Experimento materials are used the most,
- The degree of satisfaction with the materials among educators, and
- What they think could be improved.

A detailed and real-world view of classroom use comes from video observation or an observational study (evaluation section 4, chapter 8). Four didactically-oriented criteria provide guidance on how experiments can be used in classrooms to support learning:

- Do students have plenty of opportunities to actively participate in the process of acquiring knowledge?
- Are the experiments used to explore real-life problems?
- Do teachers encourage students to comment and draw upon their prior knowledge?
- Do teachers encourage reflection?

Evaluation section 5 (chapter 9) provides teachers with suggestions for assessing the effectiveness of their own lessons. The test exercises are related to three main steps in the cognitive process:

- Developing scientific questions and hypotheses
- Planning and executing scientific experiments
- Analysis and interpretation of data

A survey can assess:

- The degree of interest in science
- Children's confidence in experimenting (self-concept) and if and/or how this changes

All the instruments and processes outlined here have been tested and utilized during evaluation processes in Germany.

5 Evaluation Section 1 – How Useful are the Handouts, Student Worksheets, and Experimentation Materials?

The Siemens Stiftung education program Experimento supports early science learning with handouts for teachers, worksheets for students, and experimentation materials.

The handouts help teachers use experiments in classroom lessons effectively and in a didactically-relevant way. They are also the foundation for the teacher training seminars and serve as points of reference for the daily use of experimentation materials in classroom lessons. Student worksheets supplement classroom experiments and make it possible for students to autonomously interact with the experiments and learn at their own pace. The experimentation materials enable teachers to conduct experiments without excessive prep or planning.

Once the handouts, student worksheets, and experimentation materials are adapted for each country, the following evaluation instrument (a checklist) can ensure the quality of all materials.

Evaluation Section 1 – How Useful are the Handouts, Student Worksheets, and Experimentation Materials – contains criteria that can be used to determine if the available materials are sufficient for achieving Experimento's goals. The results provide recommendations for updating and adapting the materials.

The applicable method is a content analysis conducted by specialists or didactic experts, or by student-teachers who have been previously instructed by experts in the method. The quality of the experimentation materials is documented over the course of multiple tests and is evaluated at the end.

5.1 What defines the quality of the handouts, student worksheets, and experimentation materials?

The quality and utility of the handouts is assessed against five scientifically relevant criteria:
1. The content is part of the curriculum.
2. The handouts provide the necessary factual knowledge.
3. The handouts and student worksheets stimulate scientific thinking and methods.
4. The handouts provide didactic guidance for the effective use of experimentation in the classroom.
5. The experimentation materials are easy to use.

5.1.1 Does the content of the handouts fit within the national curriculum?

A clear connection to the national curriculum makes it easier for teachers to achieve curricular goals by using experiments in normal lessons.

Criteria for curricular compatibility: The handouts are useful if the available experiments fulfill national curricular standards and/or other applicable standards (such as PISA) in addition to covering a wide range of topics.

Assessment: Begin by entering each subject into the table (column 1) and check the compatibility of each subject against national curricular and lesson plans (preschool, primary school, secondary school).

Subject ... is part of	Preschool curriculum	Primary school curriculum	Secondary school curriculum	Other relevant national criteria

5.1.2 Does the handout contain the expertise needed to conduct the experiments?

Subject matter expertise is important to ensure that teachers feel confident and prepared for conducting experiments in lessons. The expertise helps them recognize scientific concepts in everyday situations, provide competent answers to student questions, and to competently implement experiments in lessons. Teachers, especially at the primary school level, do not usually have an academic background in science, which means the most relevant background information on a given subject should be included in the handout.

We consider two aspects when analyzing background knowledge: ‘Technical terms’ are specialist formulations that are usually not part of everyday conversation (such as ‘electrons,’ ‘amino acids,’ ‘proteins,’ ‘denaturation,’ etc.). Even if these technical terms are unlikely to be used with children, the most important terms should be identified and explained in the interest of comprehending the subject matter. ‘Concepts’ are defined as connections between two or more conditions in a substantiative (and often causal) context, (example: if a switch has interrupted a circuit, electrons can no longer flow; a lamp cannot light up in an interrupted circuit. Or: The smaller the soil components, the better the water is filtered).

Criteria for technical knowledge: The handouts are useful if they contain the necessary technical terms and concepts with accurate and easily-understood explanations.

Assessment: Enter a number or a keyword for each experiment in the table (column 1). Examine <u>every experiment</u> contained in the handout to ensure the relevant technical terms and scientific concepts are identified and easily understood.

Experiment		**Agree**	**Mostly agree**	**Mostly disagree**	**Don't agree at all**	**What's missing?**
	Technical terms are identified					
	Technical terms are easily under-stood					
	Concepts are identified					

5.1.3 Do the handouts/student worksheets contain suggestions on methods for acquiring scientific knowledge?

Scientific thinking and methods are, alongside subject matter expertise, an important element of scientific competency (Wellnitz & Mayer, 2013). There are several typical mindsets and methods that must be developed before focusing on acquiring the necessary competencies for experimentation. (Anders, Hardy, Pauen, Ramseger, Sodian & Steffensky, 2013). These include abilities such as the systematic observation of processes and facts, identifying characteristics, and measuring, categorizing, or comparing their properties. Teachers without prior knowledge of science are not necessarily familiar with the importance of this methodology, which is why the handouts and worksheets for students play an important role.

<u>Criteria for learning methods:</u> The handouts and student worksheets are most beneficial when they contain a wide variety of age-appropriate ideas for stimulating scientific thinking and methods.

Assessment: Enter the number/a keyword for the experiment in the table (column 1). Make a check next to <u>each experiment</u> if the student worksheet and teacher handout contain an explicit or paraphrased reference to the scientific method (repeat mentions are possible).

Experiment	**Student worksheet**	**Teacher handout**
	☐ Observe/reflect ☐ Compare/categorize ☐ Differentiate ☐ Measure ☐ Examine ☐ Draw/record/ document	☐ Observe/reflect ☐ Compare/categorize ☐ Differentiate ☐ Measure ☐ Examine ☐ Draw/record/ document

5.1.4 Does the handout contain the necessary expertise to conduct the experiments?

Usually, primary school teachers are not specifically trained to teach science lessons. To enable them to prepare and conduct effective science lessons anyway, the materials must support teachers with subject-specific expertise and the fundamental principles of quality science lessons (Dochy, Segers, Van den Bossche & Gijbels 2003; Morrison & Ledermann, 2003).

Criteria for subject matter expertise: The teaching methods are useful if they

a) Provide teachers with explicit guidance for problem-based lessons that focus on everyday phenomena,
b) Identify several examples for students and teacher prompts,
c) Outline prior knowledge required for students and typical misconceptions students may have, and
d) Contain specific ideas for student activities for autonomous learning and working.

Assessment a) Enter the number/a keyword for the experiment in the table (column 1). An everyday phenomenon is a situation or a circumstance that is familiar to children from their own lives. The phenomenon is suitable if it illustrates the concept of the experiment. Count the number of phenomena per experiment and check if the phenomenon is a good fit for each experiment.

Experiment		**Total number**	**Number that fit the experiment**
	How many of the everyday phenomena listed are familiar to children from their everyday lives?		

Assessment b) For each experiment, check the introductory text at the beginning of the student worksheet. Does the text present a problem or question that can be explained or answered in the experiment that follows? How many questions per experiment are included in the teacher handout and the student worksheet?

Experiment			
	Does the introduction at the beginning of the student worksheet present a problem or question that can be addressed with the experiment?	Yes	No
	How many teacher prompts are included in the teacher handout for this experiment?	Number:	
	How many questions for reflecting on the experiment are included in the student worksheet?	Number:	

Assessment c) <u>For every experiment</u>, check if the handout outlines the prior knowledge required by students and includes at least one misconception typically held by students.

Experiment			
	The necessary prior knowledge is identified.	Yes	No
	How many typical misconceptions are identified?	Number:	

Assessment d) <u>For every experiment</u>, assess the independent activities students are encouraged to do, both in the student worksheet and in the teacher handout.

Experiment		Encouraged to observe	Encouraged to ask questions	Encouraged to predict	Encouraged to plan an experiment	Encouraged to try things out
	Student worksheet					
	Teacher handout					

5.1.5 Are the necessary experimentation materials available and appropriate (deployable/user-friendly)?

The usability aspect is pragmatic: Materials that are easy to use increase the chance that they will be used. For the experimentation materials, this means it should be possible to implement the experiments quickly and safely without complicated set-up (Gichoya, 2005), that the materials are easily replaced, and that the materials are robust enough to withstand plenty of use.

Criteria for usability: The experimentation materials are useful when all the necessary materials are listed, additional materials are easy to acquire, all items are robust and long-lasting, and safety measures are sufficient:

Assessment: Two independent people should try out <u>each experiment</u> and make note of the indicators.

Experiment		Yes	No	What didn't apply? What is missing?
	All items that are needed for the experiment are listed.			
	All items that must be obtained are available around the house or are easily found in supermarkets, drug stores, or department stores.			
	All items are robust and long-lasting.			
	The safety measures mentioned in the introduction are sufficient.			

5.2 How do you assess the quality of handouts, student worksheets, and experimentation materials?

The quality of the handouts and student worksheets is determined by a quantitative content analysis (Mayring, 2015) based on the criteria described above. The tie-in to the curriculum occurs in each individual subject (thematic area), while the other criteria are assessed within each experiment.

The criteria for the material analysis were developed specifically as part of this project (Kohlhauf, Neuhaus, Haslbeck & Lankes). They can be assessed singularly or in parts.

For a reliable assessment, it is recommended that the criteria are assessed (coded) by at least two people. These individuals should be instructed in the methods until the concordance of the evaluations is acceptable (interrelated).

The usability of the experimentation materials is assessed by two or more independent observers conducting each experiment and recording their observations.

5.3 How do you interpret the results?

Observe the results of a single criterium across all subjects and/or experiments and calculate the proportion of subjects/experiments that meet the criterium. The following example includes fictitious results to demonstrate the possibilities of interpretation.

Criteria: curricular applicability – example: Experimento I 8+ contains 17 subjects. An analysis shows that three of the subjects (18 percent) are part of the national preschool curriculum, 10 subjects (59%) are part of the national primary school curriculum, and 4 subjects (23%) are not part of the national curriculum at all.

These (sample) results would show that the material in this case is best suited for primary schools – the teachers using it are adhering closely to the curriculum. It is open to discussion if the four subjects that cannot be assigned to a curriculum can be justified or should be used. If the proportion of subjects that cannot be applied to a curriculum is large, teachers can only use the materials in an extracurricular setting, such as during free periods or in working groups.

Criteria: subject matter expertise – example: Experimento I 8+ contains 43 experiments. An analysis shows that subject matter technical terms are named in 40 (93 percent) of the experiments but are only adequately explained in 34 (85 percent of 40 experiments). The applicable concepts can be recognized in all 43 (100 percent) experiments.
In this case, the teacher handouts provide most of the necessary subject matter expertise for lessons. The goal should be that the necessary expertise is identified and explained in as many experiments as possible, and that the concepts depicted in the experiments are also identified. Gaps in any of the experiments indicate room for improvement.

Criteria: mindset and working methods: In addition to the subject matter, students learn specific ways of thinking and working through experimentation. These are essential for the scientific discovery process, but they also play a crucial role in developing rational, evidence-based, and analytical ways of thinking and reasoning. These methods of thinking and working must be targeted and developed during lessons. The handouts can help encourage multiple ways of thinking and working.

The analysis indicates:

- how many thinking and working methods are present (the sum of all mentions),
- which occur frequently or infrequently (proportion of the number in percent),
- if the thinking and working methods are identified predominantly in the teacher handout (teacher-induced) or in the student worksheets (self-induced),
- which experiments – or collectively: which subjects – touch upon thinking and working methods frequently or infrequently. This would be another indication of the need to develop materials further.

Criteria: subject matter didactics: The four aspects and the related questions are evaluated for each experiment. For each experiment, at least one everyday phenomenon from the daily lives of children should be identified. An experiment's introduction should reference a problem or a question that can be at least partially explained by conducting the experiment. The teacher handout for each experiment should contain several questions that serve as learning impulses. The student worksheet for each experiment should contain several questions that encourage children to think critically and explore. For each experiment, but at the very least for each subject, the teacher handout should also contain the prior knowledge needed by students and at least one anticipated student misconception. The more stimuli for activating the students in the teacher handout and for independent learning in the student worksheet, the better.

Criteria usability: Ideally, each experiment will have as many questions as possible answered with "yes." Questions that are answered with "no" are often an indicator of the need for revision.

6 Evaluation Section 2 – Quality of Teacher Training

Experimento provides teachers with a selection of materials relating to energy, the environment, and health that are suitable for everyday classroom use and aligned with the curriculum. Teacher training seminars demonstrate how to conduct the experiments and use them in lessons.

To embed scientific experiments practically and effectively in learning situations (Corrigan, Dillon & Gunstone, 2011), teachers need subject matter expertise. Nationally-coordinated seminars for teachers are part of Experimento. The seminars develop teachers' expertise and confidence until they can implement the materials as envisioned by the stated goals of Experimento (Van Veen, Zwart & Meirink, 2012).

Domain-specific knowledge, in this case action-guiding scientific knowledge, is divided into technical and didactic knowledge. However, it is not enough to be able to factually draw upon this knowledge. The related attitudes and confidence are also essential for putting this knowledge to use. A fundamental element of the learning process is that teachers are satisfied with their training and that they embrace it.

We assess the quality of teacher training in four aspects:

1. Does the training provide the necessary technical expertise that teachers need to competently implement the experiments?
2. Does the training provide the necessary didactic expertise for teachers to teach content and methods in an understandable way?
3. Do teachers have the opportunity to practice subject-specific and didactic expertise during training to gain confidence in using experiments?
4. How satisfied are teachers with the teacher training?

Evaluation Section 2 – Quality of Teacher Training – provides ideas on how the impact of training can be assessed based on the technical and didactic expertise provided and the safe and confident use of experiments in lessons. The results can serve as indicators for potential revisions to training content.

6.1 What do we expect from training seminars on scientific experimentation?

Establishing action-guiding knowledge is a long-term process that is influenced by several factors. Training seminars are usually limited in scope, making their permanence and long-term effectiveness difficult to assess. (Darling-Hammond, Hyler & Gardner, 2017). At the same time, we expect demonstrable growth in knowledge and confidence directly following a seminar.

6.1.1 Methodical expertise

When it comes to scientific knowledge, we differentiate between subject matter expertise about scientific terms, facts, and concepts, and familiarity with typical scientific methods for generating and evaluating subject matter expertise (Shulman, 1987).

Even when preschool or primary school children are only learning basic, age-appropriate content and methodology, teachers need a deeper understanding of facts and concepts in addition to a broader understanding of the related methods. Subject matter expertise has proven to be a relevant contributor to didactic ability, meaning a person lacking solid understanding of the subject matter has trouble explaining even simple scientific principles.

Subject matter content (terms, facts, concepts) is specific for each experiment. As is the case with the teacher handouts, we also expect teacher training to identify the necessary subject matter terminology and concepts, in addition to explaining these rationally and accurately. However, because a given seminar can only prepare and execute a certain (unspecific) number of potentially available experiments, each seminar will cover different sets of subject matter expertise. This makes a general assessment of the subject matter expertise as part of an evaluation not possible.

But each experiment implemented in the classroom utilizes the same methodological process: At the beginning, a hypothesis is proposed, and an experiment is planned that is suitable for testing the hypothesis. During or after the experiment, the results are observed, recorded, and evaluated against the previously proposed hypothesis. This is known as the scientific method.

To correctly prepare the specifics of an experiment and to apply the scientific method with children, teachers must themselves be able to:

- Formulate a hypothesis
- Plan an experiment
- Evaluate data

These three competencies are assessed in a test on methodical expertise, which can assess if teachers' methodical expertise increases because of the seminars.

6.1.2 Didactic expertise

Teachers need didactic expertise to convey the subject matter and scientific method to children in an easily understandable way (Shulman, 1986, 1987). Didactic expertise influences students' classroom learning outcomes (Förtsch et al., 2018) and is a decisive factor for the quality of lessons. A didactically competent teacher is aware of typical difficulties and misconceptions of students and can draw upon best strategies for using experimentation to create learning opportunities.

To offer experiment-based lessons that maximize student learning and utilize their prior knowledge, the seminars familiarize teachers with:

- Student misconceptions and how to support learning when they make mistakes,
- Information about instruction and teaching strategies.

These two areas are combined in a test on didactic expertise and – because didactic expertise can be applied and assessed outside of a subject matter context – assessed on one primary school-level lesson in biology (respiration) and one in physics (electricity). The test can determine if teachers' didactic knowledge increases due to the training seminars.

6.1.3 Gaining confidence in experimentation

Beyond expertise and abilities, personal conviction and attitudes – such as a person's sense of their own abilities (self-concept) – also influence the willingness of teachers to use experiments in the classroom (Klingebiel & Klieme, 2016): A person who feels competent and confident in a certain field is more likely to engage in or take on the task of experimentation. A person who is fearful of a subject, or who feels uncertain or insufficiently prepared, is more likely to avoid such subjects or tasks. This also applies to teaching both technical and methodical subject matter. If teachers have the opportunity during training to apply technical and didactic expertise on examples, we can expect that their self-concept increases, including their confidence in experimentation.

6.1.4 Acceptance and satisfaction with the training

One important factor for successful teacher training is that the participants are satisfied with the seminar and its content (Lipowsky, 2010). Satisfaction increases when the subject matter is perceived to be interesting, fit for purpose, and relevant. Additionally, satisfaction increases when instructors are relatable and open with participants and when the seminar is conducted in a pleasant environment (Smith & Gillespie, 2007; Lipowsky, 2010).

We assess teachers' satisfaction:

- With the subject matter of the seminar,
- With the instructors,
- With the environment.

After the training, we use a survey to assess if the teachers were satisfied with the content, the environment, and the instructors.

6.2 How is the impact of the seminar assessed?

We use a test to assess learning growth in methodic and didactic expertise. Developing a reliable test is a complex task that requires knowledge of methods and experience. Therefore, a better approach is to utilize existing tests, such as the established Experimento test for teachers described below.

Teachers' existing knowledge is determined in a test that is administrated before the start of the teacher training seminar. The training should take this prior knowledge into consideration and build upon it. After the training, the same test shows the learning progress of the participants.

Self-concept is evaluated using a questionnaire before and after the seminar. Satisfaction with the training is assessed with a questionnaire after the conclusion of the training.

To allow for the pre- and post-tests to be collated and compared, participants are asked to create and include a unique identifier code on each test.

6.2.1 The Experimento test on methodic expertise

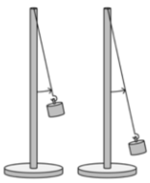
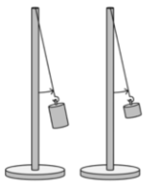
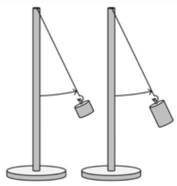
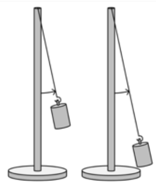
The Experimento test of methodic expertise draws upon proven exercises from existing studies (Vorholzer, 2016; Arnold, 2015). In some cases, these have been modified and amended with additional exercises.

The exercises pertaining to the partial construct of ‘forming a hypothesis’ target the alignment of experimental setup and questions/hypotheses. They examine the understanding of the function of a hypothesis.

The systematic development of an experiment is dependent on the ability to distinguish between independent and dependent variables. To establish a causal relationship of an (independent) indicator and a (dependent) variation, only one independent variable can change; all other independent variables or confounding variables must remain the same or be controlled (variable control strategy). The exercises pertaining to the partial construct ‘planning experiments’ target the recognition of the independent and dependent variables in addition to confounding variables. They also examine the understanding of variations to an independent variable.

Each evaluation begins with documentation of observations. This must precisely distinguish between observation and (the resulting) interpretation before any conclusions can be determined. This is examined by the exercises pertaining to the partial construct ‘data evaluation.’

Markus has planned a series of experiments for examining the oscillations of a thread pendulum. He is considering the following four experiments, which are all effective in examining the influence of a single variable.

Experiment A	Experiment B	Experiment C	Experiment D
			

Which of the experiments listed above do you consider the best for examining the question: Is the period of oscillation affected by the length of the thread?
Please check only <u>one</u> answer.

- ☐ Only experiment A
- ☐ Experiments A and D
- ☐ Experiments A, B and D
- ☐ Experiment D
- ☐ None of the experiments

Sample exercise on methodic knowledge, partial construct ‘planning experiments’ (based on Vorholzer, 2016)

6.2.2 The Experimento test on didactic expertise

The Experimento test on didactic expertise was developed as part of this evaluation project (by Kotzebue, Haslbeck & Neuhaus). The test is based on preliminary work by Jüttner, Boone, Park and Neuhaus (2013) and by Kotzebue and Nerdel (2012).

The exercises examine teachers' didactic expertise based on the examples of electricity (physics) and respiration (biology) in four partial constructs:

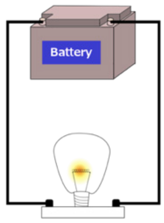
- Knowledge of student misconceptions and the correct way to address mistakes,
- Knowledge of instruction and learning strategies,
- Knowledge of the use of models as a visualization strategy, and
- Knowledge of how scientific concepts can be illustrated and understood using everyday phenomena.

Mathilda and Leon want to find out if the size of a battery influences the brightness of a light bulb. They conduct an experiment:

Condition 1:

Room temperature 28°C

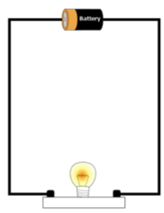
A large battery, a large bulb, and two wires



Condition 2:

Room temperature 10°C

A small battery, a small bulb, and two wires



The large bulb is brighter. The children conclude that the battery impacts the brightness of the bulb.

Student misconception: Mathilda and Leon do not notice that they cannot come to any causal conclusions in this experiment because three factors (battery, room temperature, and the bulb) were changed.

How can you as a teacher react to this kind of mistake in the classroom?
Please list three ideas.

____

____

____

Sample exercise on didactic expertise, partial construct 'student misconceptions and how to address mistakes' (based on Kotzebue, Haslbeck & Neuhaus)

6.2.3 The didactic self-concept questionnaire

The questionnaire on didactic self-concept uses questions from existing studies.
It contains:

- A general scale for estimating a person's own abilities for using experiments in lessons (Kuhn, Lankes & Steffensky, 2012),
- A scale for estimating a person's own abilities on lessons about electricity (Böttcher-Graf, 2016),
- A scale for estimating a person's own abilities on lessons about respiration (Böttcher-Graf, 2016),
- A scale for estimating a person's own abilities on explaining the variable control strategy (Haslbeck, 2019).

Scale	Sample item
Didactic self-concept on the use of experiments	I can explain scientific phenomena at the level of a child
Didactic self-concept electricity	I feel confident integrating experiments for students about 'electricity' into lessons
Didactic self-concept respiration	I feel confident integrating experiments for students about 'respiration' into lessons
Didactic self-concept on variable control strategy	I can explain to children what happens when more than one variable is changed in an experiment

Sample items from the scales on didactic self-concept
(Kuhn, Lankes & Steffensky, 2012; Böttcher-Graf, 2016; Haslbeck, 2019)

6.2.4 The satisfaction questionnaire

The satisfaction survey contains a scale on satisfaction with the subject matter, a scale on satisfaction with the instructors, and two single items: one item on the environment and one item on overall satisfaction.

Scale	Sample item
Satisfaction with the material	I feel well-prepared to implement Experimento in the future.
Satisfaction with the instructors	The instructors presented the material effectively.

Sample items from the scales on satisfaction.

6.3 How do you interpret the results?

The three instruments – the tests on methodic and didactic expertise in addition to the survey on didactic self-concept – should, if possible, be administered before and after the seminar to demonstrate its effectiveness.

Prior to the training, the three instruments provide training organizers with important information about the prior knowledge of the participants. Organizers gain insights into areas such as:

- How familiar are the participants with the scientific process? Are they able to formulate verifiable hypotheses, to systematically plan an experiment, and to draw reasonable conclusions from the data?
- Do teachers know the difference between student perceptions that are correct, potential learning opportunities, or false? And can the teachers use experiments to help students realize for themselves that their perception is incorrect?
- Can teachers use visualization models and are they familiar with phenomena that children would know from their everyday lives that would help explain concepts?
- How confident do teachers feel when using experiments in the classroom? Are there differences in their self-concept of biology subjects (such as respiration) and physics subjects (such as electricity)?

The organizers of the training seminars can build upon the prior knowledge of the participants, going into more depth or expanding coverage of certain areas as needed.

After the training seminar, the instruments show if the goals of the training have been achieved: if teachers have expanded their methodic and didactic expertise and have gained confidence in using experiments in lessons.

The survey on satisfaction provides general feedback on participants' experience in the seminar.

7 Evaluation Section 3 – How do Teachers Use Experimento?

Experimento aims to encourage and support early science learning. Providing materials for the range of experiments requires personnel and financial resources, which is why it is important to ensure that the materials are put to good use and that they fulfill their purpose.

Evaluation Section 3 on teachers' use of Experimento provides insights into what parts of the program are used often, which are used rarely, and which are never used at all. Feedback also shows how satisfied the teachers are with individual experiments and areas they would like to see improved. The findings pertaining to the didactic use of the training seminar and handouts is important – this demonstrates teachers' personal assessment of how they use the materials and if it is in line with how implementation is envisioned by the program.

7.1 How do we expect the material to be used?

We address the following questions from a cost-benefit perspective:

1. How often are the materials actually used?
2. How do teachers assess the usability of the materials?
3. According to teachers, is the use of the materials in the classroom in line with the intended didactic principles? (See also chapter 2.5)

A logbook is used to explore these questions.

7.1.1 Frequency of use

Experimento materials that are collecting dust in a drawer equate to wasted resources. To ensure that the materials fulfill their purpose and achieve stated goals, they must be put to use.

The experimentation materials should be used for as many experiments as possible at a given school. Guidance from the handouts and the student worksheets are valuable didactic aids that should be used. Experimento should become a permanently-established principle in science lessons at a school, which is why as many teachers as possible should use it.

We expect that the highest number of different experiments will be used by as many teachers, in as many classes as possible.

7.1.2 Usability as a condition for implementation

The acceptance, the use, and the success of the materials depends on how these are perceived by the people using them (Stergioulas et al., 2014; Venkatesh et al., 2003). 'Usability' is one commonly-used indicator to measure these aspects (Nielsen, 2000).

The globally-recognized norm for creating systems that are particularly fit for purpose defines usability as the degree to which a product can be used by a certain user in a certain use case to effectively, efficiently, and satisfactorily achieve a certain outcome (DIN EN ISO 9241-11, 2016). This definition can be used to assess the perception of usability for products and materials in the education sector.

In this context, high usability is achieved if the materials are complete and long-lasting, and if the information is easily understood. If materials must be acquired, they should be easy to come by. Making Experimento materials quickly and easily accessible for all educators is another relevant factor affecting usability.

7.1.3 Didactic use

The Siemens Stiftung education program Experimento seeks to maintain and expand children's interest in science subjects and to support early science learning. Teachers use inquiry-based learning to actively involve children in the learning process while helping them form their own views of the world (see also chapter 2.5).

This is done by teachers who give children multiple opportunities to explore the material on their own, come up with their own questions and hypotheses, plan and conduct experiments, record their observations as they see fit, and interpret results on their own.

7.2 How is usability assessed?

Reliable information about usability is best obtained with a logbook. A logbook collects information about the use of materials alongside the process of implementing them. In short surveys, teachers document the experiments they have conducted and how, in addition to how satisfied they were with the results. The Experimento logbook was developed as part of this evaluation project (Haslbeck & Lankes). It should be kept in the same place as the experimentation materials.

One challenge of the evaluation is convincing teachers to fill out the survey after each and every use of the experimentation materials. The survey is deliberately short and can be completed in about five minutes; at the conclusion of an experiment when the children are putting away the materials is a good time. The teacher training seminar should always emphasize the relevance of the logbook, and teachers should be willing or even compelled to use it. Arguments for using the logbook include:

- The information helps us adapt the experimentation materials based on how they are actually used, and to continuously improve the handouts.
- The information helps us to target teacher training to the specific application of Experimento in lessons.

Guaranteeing anonymity for the collection and evaluation of data increases the reliability of the answers. Therefore, the data should not enable any connections to be drawn to a particular school or individual teachers (anonymous submission, no indication of school in the cover sheet or other documentation).

After the introductory text, the logbook contains a page that is filled out by the school director (cover page).
This includes information about the size of the school, the number of teachers and classes, and the significance and scope of Experimento deployment in the entire school.

At the end, the logbook contains two pages of questions for documenting each experiment. Each experiment survey begins with the date and the class name.
Questions 1-4 ask which experiment or experimentation materials were used.
Questions 5 and 6 provide brief insights into the didactic methods used.
Questions 7-10 allow conclusions to be drawn about perceived usability.
Questions 11 and 12 are about teacher satisfaction.

Page 3 and 4 of the logbook (repeat for 20 experiments)

Experiment questionnaire (to be completed by the teacher after each experiment)

Experimento ☐ 4+ ☐ 8+ ☐ 10+ Date: ___________ Grade level: ___________

I. You conducted an experiment during your class. Which of the following did you do?

1) I used an experiment from the guide book:
☐ yes, number: __________ (continue to question 4)
☐ no (continue to question 2)

2) I used the following experimentation materials:

____

3) I used the materials to conduct the following experiment/explore the following scientific concept:

____

4) I used the student worksheet:
☐ yes, unaltered ☐ yes, altered ☐ no

II. How did you use the materials during lessons?

5) Was the experiment demonstrated for the class?
☐ yes, by me as the teacher
☐ yes, by a student
☐ no

6) Which steps were the students able to complete themselves (alone or with a partner/group?) (Multiple answers possible)

The students...
☐ established their own question.
☐ developed their own hypothesis.
☐ were allowed to examine/test the materials as they saw fit.
☐ planned the steps of the experiment on their own.
☐ conducted the experiment alone.
☐ conducted the experiment with a partner or in a group.
☐ made their own decision on how to present/depict the results.
☐ made their own interpretations of the results.

III. Were the materials complete and readily available?

	Yes	No
The guide book contained complete and easily-understood information needed for this experiment.		
All items listed for this experiment were included in the experimentation materials.		
Any additional materials were easy to obtain.		
All materials were still intact after they were used.		

IV. Were you satisfied with the materials?

I will conduct this experiment again.
☐ yes ☐ no, because:

____

____

What could be improved?

____

____

Logbook pages to be filled out by teachers following the implementation of an experiment.

7.3 How do you interpret the results?

If they are consistently completed and submitted, logbooks provide valuable and comprehensive insights into classroom implementation. The results can be used by Experimento program coordinators and organizers to update training and/or the materials. The more completed logbooks that are submitted, the more reliable the data becomes as it is aggregated and evaluated.

7.3.1 Indicators about frequency of use

- The school director indicates how many teachers, from which class, and in which grade level have taken part in the Experimento teacher training. The documentation of the experiments indicates how much time has passed between the seminar and the first use of Experimento for lessons. The best case would be for the first use of Experimento to come as quickly as possible after the end of the seminar, so the learned material is still fresh.
- Teachers indicate which experiments or which experimentation materials they used and if they made use of the student worksheet. It is useful to number or tag the experiments so the data entered can be distinctly categorized. Taken collectively, the input from several logbooks provides an overview of the materials, handouts, and/or school worksheets that are used more or less frequently. If something is rarely used, it is worth asking why. Depending on the reason, that particular asset can either be set aside (for example, if it doesn't fit the curriculum) or improved upon (if it was not robust enough or easy enough to understand).

7.3.2 Indicators about usability

- A school director decides where the Experimento materials should be stored and if this room is accessible to all teachers. Uncomplicated access is an important contributing factor toward putting the materials to use. For example, keeping the materials in one classroom would reduce other teachers' ability to use the materials flexibly and as needed.
- After each experiment, teachers indicate if the information was easy to understand, if the listed experimentation materials were available, if materials that needed to be obtained were easy to find, and if all materials were still functional after the experiment (assuming they were not meant to be single use). Aggregated across several experiments, this information provides insights into potential areas for improvement. Questions pertaining to repeated use create possibilities for exploring additional criticisms or desired improvements.

7.3.3 Didactic indicators

- Question 5 indicates if, and if so, by whom, an experiment was conducted. A teacher or a student conducting an experiment on their own is not in line with the didactic principles of the Experimento program. Every student should conduct the experiment as part of a group or with at least one partner.
- Question 5 also gives insights into the active participation of students in the discovery process. Ideally, students become progressively capable of autonomously engaging in the discovery process. Initially, they need more direction and support, with eventual next steps successively added and practiced. In higher grade levels, every student should be capable of creating a research question based on an observation (such as an unexpected result), creating and testing predictions based on prior knowledge and formulating measurable hypotheses, planning an experiment, documenting observations, and drawing conclusions. To arrive at that point, students must be given a degree of freedom and be driven by their own personal initiative, which will also grow over time.
- School directors determine the teachers who take part in the teacher training. The documentation of the experiments shows if only the teachers who have received specific training are the ones using the materials or if the school manages to create a new culture of science lessons that extends to all classes.

8 Evaluation Section 4 – Quality Lessons with Experimento

Experimento seeks to equip teachers with the ability to encourage and support early science learning. This is the goal for which the experimentation materials have been created and the handouts and worksheets developed. In a blended learning seminar, teachers learn how they can use the Experimento materials as envisioned by the program: implementing experiments in science learning situations in a way that is easy to understand and supportive of permanent learning.

Evaluation Section 4 – Quality Lessons with Experimento – provides ideas for examining the long-term impact of the teacher training seminars on practical implication, meaning the way in which teachers use the Experimento materials in their lessons. A reliable assessment method is classroom observation, either through a participatory observer or in a video analysis. The findings provide insights into potential revisions to the training and the teacher handouts.

The instruments can also be used for certain types of self-evaluation or in feedback sessions, such as sitting in on peer classes. They show teachers what is most relevant when it comes to creating learning opportunities for science by providing ideas and tips.

8.1 How should experiments be used in lessons?

From a science didactics perspective, cognitive activation is the key indicator of good lessons and/or effective learning situations (Steffensky & Neuhaus, 2018). Cognitive activation is also among the basic dimensions of lesson quality, alongside classroom management and constructive support.

Cognitive activation measures are therefore among Siemens Stiftung's main goals for encouraging children to conduct experiments on their own while coming up with questions or reflecting on solutions. (Osborne, 2014; Kotzebue, Müller, Haslbeck, Neuhaus & Lankes, 2020).

For observing the use of experiments in lessons, we have specified four central aspects of cognitive activation. While they also apply in other contexts, here they are specific to science lessons:

1. Activity-oriented
2. Problem-oriented
3. Active utilization of students' prior knowledge and perceptions
4. Encouragement for students to engage in further reflection

Beyond these, general indicators of good lessons that apply in every subject include a clear, goal-focused structure, or the previously mentioned basic elements of classroom management and constructive support.

8.1.1 Activity-based

In line with the principle of inquiry-based learning, teachers can ensure that children are active participants in their own learning process. Teachers guide children as they form their own view of the world (see also chapter 2.5 and 7.1.3). This can be achieved by giving children many options to make their own progress in the discovery process. Examples of this include the freedom to learn about materials on their own, forming their own questions and assumptions, planning and executing their own experiments, depicting their observations how they see fit, and interpreting these on their own.

As with the logbook (See also Chapter 7, Evaluation Section 3) we can initially use classroom observation to see who conducted the experiment:

- The teacher demonstrated the experiment to the class.
- One student demonstrated the experiment to the class.
- The students conducted the experiment in groups or with a partner.
- Each student conducted the experiment alone.

The second step documents which steps of the cognitive process are actively and independently conducted by the children:

- Freely exploring the materials
- Developing research questions
- Forming predictions
- Planning the steps of the experiment
- Documenting/protocolling/recording observations
- Interpreting results

Finally, action-oriented means practicing and implementing various scientific mindsets and methods, which – with input from teachers, if needed – is done by the students on their own. We consider the following fundamental scientific mindsets and methods to be age-appropriate:

- Observation/consideration
- Comparing/categorizing
- Variation
- Measuring
- Examining

We expect students to be capable of independently executing as many of these steps as possible, either alone, with a partner, or in a group.

8.1.2 Problem-based

The inquisitiveness, everyday problems, or personal observations of children establish a willingness at the beginning of a learning situation to engage in the activities described below by creating curiosity, raising questions, and stimulating mental activity (Kobarg & Seidel, 2007). Successful teachers use exciting, challenging problems to prepare learners for the experiment, to inspire questions, and to incorporate personal experiences and assumptions. Over the course of a lesson, children come back to the original problem and can answer the initial question based on the results of the experiment.
Not only is this good motivation for learning, but it also activates mental processes, linking new knowledge with prior knowledge and stimulating deeper learning processes.

We expect that teachers will introduce the experiment at the beginning of a lesson through a motivating problem or phenomenon and use this problem to allow questions or predictions to develop. These can be answered during the lesson based on the results of the experiment.

Scale	Example item
Problem-based	An initially-raised question/problem is revisited by the teacher throughout the lesson (through exercises/experiments, repeated paths that link back to the problem, etc.). Applicable – mostly applicable – mostly not applicable – not applicable

Sample item for the problem-based scale.

8.1.3 Encouraging student participation and activating prior knowledge or predictions

Language is one indicator of cognitive activity that is also itself a cognitive activity. Language development is among the central purposes of early childhood education. Experimento encourages teachers to use science learning situations to support children in the development of their language abilities. Examples of how this is done include allowing children to formulate their own questions, vocalize their predictions, use technical terms during experiments, describe phenomena, discuss problem-solving ideas, and draw conclusions.

We expect that students are given many different opportunities to express themselves vocally (including with their partners or groups) over the course of a lesson.

Scale	Example item
Encouraging student participation and activating prior knowledge or predictions	The teacher asks about learners' prior knowledge and/or encourages the learners to make predictions.

Sample item for the scale encouraging student participation and activating prior knowledge or predictions

8.1.4 Inspiring reflection

Teachers encourage children to reflect and provide them with opportunities to discuss problems, potential solutions, misconceptions, and ideas. Unexpected phenomena, surprising results, misconceptions, conflicting explanations, and similar outcomes can create a cognitive conflict for students that leads to reflection (cognitive activity). Children are encouraged to further engage with a learning topic and to develop their knowledge.

We expect teachers to provide many ways of stimulating and motivating children to reflect.

Scale	Example item
Inspiring reflection	The teacher confronts learners with situations, observations, or phenomena that are contrary to the expectations of the learners and/or draw their attention to conflicts/inconsistencies.

Sample item for the scale 'inspiring reflection.'

8.2 How can quality of the implementation be observed?

In contrast to a teacher survey about their lessons, observation provides us with a much more precise data set that pertains to a specific situation. This can occur in two ways:

A video analysis provides the most comprehensive information about how the lesson unfolds, especially if several cameras are used to capture the teacher and the students. After the lesson, the videos can be assessed at leisure using special analytical software (see examples, including Seidel, Prenzel & Kobarg, 2005).

However, this type of analysis has some disadvantages. It is costly and human-resource intensive. The complexity and the volume of information makes the analysis more difficult. It is difficult to convince teachers to let themselves be filmed. Usually, data protection considerations mean permission needs to be obtained from the parents/guardians of children who are to be filmed.

At the same time, a video captures and retains all relevant information, and it is possible to analyze the tape, note certain situations, and compare occurrences and timing. This can be done multiple times and against a large set of criteria, if needed. Evaluations by two people (double coding) can ensure conformity between the evaluations (interrelated), thereby increasing the validity of the results.
The scales for the examples mentioned above, pertaining to the analysis of lesson videos/ observations, were developed based on a catalog of criteria from Ewerhardy (2010), Förtsch, Werner, Dorfner, v. Kotzebue and Neuhaus (2018), Vehmeyer (2009), Lotz, Lipowsky and Faust (2013), and Möller (2016)

A participatory observation requires fewer resources and less effort. The indicators to be observed are relatively easy to code, such as who conducted an experiment or which mindsets and methods can be observed. More difficult indicators are those that occur throughout the entire lesson (for example, 'teachers provide learners with time and opportunities to ask their own questions'). One option is to keep a tally to at least track the frequency of these occurrences.

However, evaluating the quality of the occurrence (rating) requires a degree of competence and experience. Ideally, observers should therefore be familiar with the material (as a teacher, university student, etc.) and receive prior training. All the same, monitoring many aspects at the same time can be difficult for a single observer. A focused selection of several clearly observable indicators makes the task more manageable.

If observations are taking place as part of sitting in on a colleague's lessons, agreeing in advance on a few specific indicators is an option. In that case, a feedback session can take place right after the lesson.

8.3 How do you interpret the results?

Lessons are a very personal, situational, and always unique event. The three central influencing variables – students and the class, subject matter, and the teacher – are in a constantly shifting relation to each other, making it difficult to make blanket statements about what objectively makes a good lesson. Therefore, it is important that lesson observation, whether participatory or by video, focuses on specific, distinct, expectable, and visible indicators.

If, for example, it is repeatedly indicated in the teacher training and in the handouts that children should be encouraged to formulate their own questions and express their predictions, then we would expect that this would take place in lessons; we can observe if this is the case, and if so, how often it occurs. To interpret the data, we then need comparison values or expected values.

If a large amount of data exists, we can make social comparisons: What is the median frequency of the occurrence across all cases? What is the frequency distribution? How large is the group of those who do not demonstrate the expected indicator at all? Results from such analyses are an important factor in improving teacher training and handouts.

If the sample sizes are too small to deliver meaningful median values, expected values based on solid theoretical principles can also be used as criteria. These are based on what would be expected to be a standard value (or a minimum) for a given criterion. For example, one expectation could be that for each experiment, at least one phenomenon from the everyday lives of children is identified.

When an observation sheet (chapter 8.2) for self-evaluation or peer consultations is used, we do not assess the quality of the lessons in general. Instead, we observe indicators – agreed upon in advance, in certain cases – and provide feedback on these indicators.

9 Evaluation Section 5 – What do Students Learn from Scientific Experimentation?

All measures taken in an education system – from investments in lesson materials or books, teacher training and continuing education, lesson duration, funding initiatives, or school facilities – need to impact students and should provide support for teachers and learners.

The education program Experimento seeks to support global STEM education. Therefore, we expect that students' STEM abilities and knowledge would increase by using Experimento.

However, we feel it is appropriate to caution against unrealistic expectations for student performance tests, such as before and after evaluations. Usually, it is only research that makes use of highly-standardized tests that can prove the impact of a particular curriculum on the performance of students. The most impactful curricula in this regard also tend to be narrowly defined. For such a study to work, all influencing factors must be precisely recorded and controlled.

Experimento was designed as a teacher training offering and not as a scientific study. Therefore, many relevant factors cannot and should not be controlled in most cases, including:

- The frequency of Experimento implementation over a certain period
- The selection of experiments that are used
- Students' prior knowledge
- Parallel implementation of other lesson plans, teacher trainings, course books, or materials
- Engagement with scientific subjects outside of school, etc.

This complicates a valid examination of the generalized effects of Experimento – from teacher training and the use of the materials in lessons to student performance.

Evaluation Section 5 – What do Students Learn from Scientific Experimentation? – provides instruments for teachers to use in their own classrooms and/or before and after specific units to assess the students' starting points and note progress and change. However, Experimento does not exclusively focus on cognitive abilities. Motivational factors such as children's interest in science and confidence in conducting experiments should also be addressed.

9.1 What should students learn by working with Experimento?

Science education is based on the education concept of scientific literacy, which considers basic science education as a multidimensional construct (Bybee, 2015). In addition to the acquisition of knowledge about scientific concepts and interrelationships (scientific expertise), knowledge about scientific mindsets and methods (expertise in the field of science) and the development of interests and motivation should be supported through early exposure to science learning. (Müller, Prenzel, Seidel, Schiepe-Tiska & Kjaernsli, 2016; Schiepe-Tiska, Roczen, Müller, Prenzel & Osborne, 2016).

9.1.1 The ability to experiment

The importance of early science learning is a guiding principle behind the Siemens Stiftung education program Experimento. The program seeks to support children and adolescents in autonomously exploring, reflecting upon, and understanding science and technology issues, in addition to developing methods and strategies for gaining knowledge through experimentation.

The program emphasizes experimentation, which is why these methods of gaining knowledge were selected for developing tests. Despite the differences inherent in various models of scientific thinking as a problem-solving process, they all contain three main components that are differentiated here (for an overview, see examples in Nehring, 2014; Vorholzer, 2016):

- Formulating scientific questions and hypotheses:
  - Distinguishing between verifiable and unverifiable hypotheses, and
  - Formulating individual predictions/hypotheses in response to a given question.
- Planning and conducting scientific experiments:
  - Selecting an appropriate experiment for a particular question/prediction,
  - Planning suitable experiments autonomously with consideration to the variable control strategy, and
  - Justifying the decision
- Assessing and interpreting data:
  - Distinguishing between valid and invalid interpretations, and
  - Autonomously assessing results.

9.1.2 Motivational conditions

Current teaching and learning research examines interest and a positive self-concept as motivational conditions for learning (such as Schiepe-Tiska et al., 2016). Personal interest provides motivation for the sustained exploration of a subject, encourages deeper and permanent learning, and creates a sense of ownership and identification (Krapp, 1992).

Academic self-concept describes a person's attitude toward their own scholastic aptitude. Actions and motivations are mutually influential, which means scholastic aptitude impacts a person's self-concept in the same way self-concept impacts aptitude development (Ehm, 2012; Guay, Marsh & Boivin, 2003; Marsh & Craven, 2006).

The goal of Experimento is to build up permanent and sustained interest in science subjects and to develop a positive self-concept when it comes to science (Osborne, Simon & Collins, 2003; Tytler & Osborne, 2012).

9.2 How should teachers assess what students have learned?

9.2.1 The Experimento test on experiment competency

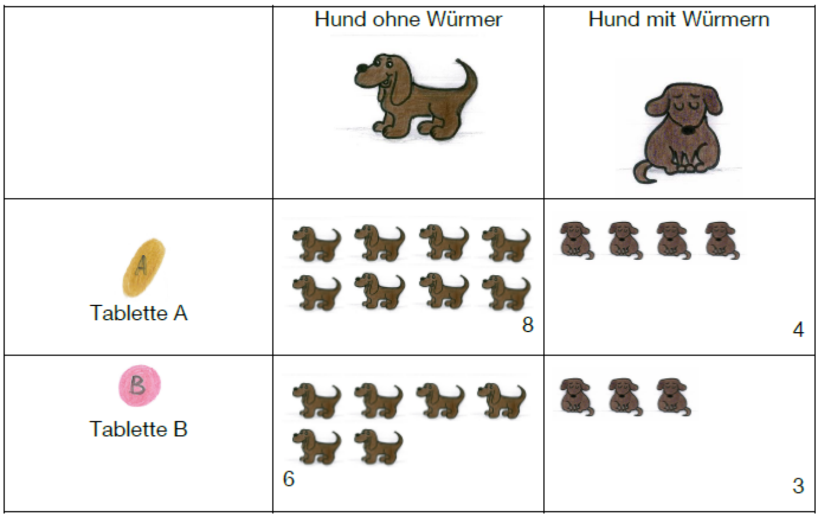
Tried and tested exercises from existing studies have been incorporated into the Experimento test examining the experiment competency of students (Saffran, 2016; Schwichow, 2016). In some cases, these have been adapted or supplemented with new exercises.

Exercises for ‘Development of scientific questions and hypotheses’ examine if students can create their own predictions/hypotheses for a particular question and/or can differentiate between verifiable and unverifiable hypotheses.

Exercises for ‘Planning and executing scientific experiments’ examine if students can select an experiment that is suitable for their problem/prediction from a range of potential experiments, if they can select a suitable experiment that takes the variable control strategy into consideration, and if they can justify their decision.

Exercises on ‘Assessing and interpreting data’ examine if students can distinguish between valid and invalid interpretations and if they can evaluate their own results.

Mr Krause tested Pill A and Pill B on dogs with worms.
The table shows how many dogs no longer had worms after taking each pill.



Examine the results closely!
Is Pill A or Pill B more effective, or are they equally effective?
Select the correct answer:

☐ Pill A
☐ Pill B
☐ equally effective

Sample exercise on the ability to experiment, partial construct ‘data assessment’ (Saffran, 2016)

Due to the limited number of items, the various subsections cannot be statistically separated from each other (this is also described in other studies, such as Wellnitz & Mayer, 2013). Therefore, a comprehensive scale is applied that extends beyond the subsections. Assessments of single item levels are also possible.

9.2.2 Survey on students’ interest in science and their self-concept

The survey on students’ interest in science and their self-concept in science subjects contains two scales. The scales were selected from a study by Kauertz, Kleickmann et al. (2011)

Scale	Example item
Interest in science	“I want to learn more about science subjects.” 1 = do not agree, 2 = agree a little, 3 = agree a lot, 4 = agree completely
Science self-concept	Understanding science is easy for me.

Sample items from the scale ‘interest and self-concept.’

9.3 How do you interpret the results?

The test on the ability to conduct experiments can be given by teachers at the beginning of a school year to provide a general impression of students' existing knowledge. This does not require any special expertise. If the test is given before or after a specific unit, it can show if students have improved their abilities. In this case, it is important to note that short units of just a few hours will only be able to achieve a short-term learning outcome, if one is achieved at all. Permanent competency development requires long-term, repeated, and successively expansive learning opportunities. In addition to average test scores of each student, results on single items are also relevant for teachers. They show strengths and weaknesses of an individual child with more precision in addition to demonstrating where individual support is most needed in lessons.

Interest is a construct that develops slowly through repeated and positive experiences and interactions with a subject. The questions pertaining to interest should therefore be repeated after a longer amount of time has passed. If lessons pursue the goal that all students should have a certain interest in science subjects, the survey can provide indicators about the motivational requirements of the students when it is administered at the beginning of the school year. Building interest can be supported in several ways, such as experiences of success; interesting, age-appropriate, and relevant problems and questions; or the opportunity to independently try things out and discover.

The survey on science self-concept can provide teachers with important information about the learning situation of individual students and is a good place to start for gaining an initial impression at the beginning of the school year. Administered at the end of long units or at the end of the school year, the survey can show if students' perception of their own abilities has changed.

10 Conclusion and outlook

You have implemented the education program Experimento because you want to bring about change: the scientific abilities of educators, the learning opportunities in science and lesson outcomes, student learning and their attitudes, abilities, and competencies.

Programs aimed at improving student performance are complex, and they cost lots of money, time, and energy. Anyone hoping to tailor a particular measure to a group of stakeholders, the requirements of teachers, and the abilities of students, shouldn't leave the impact of these measures to chance.

Educational evaluations go beyond simply assessing outcomes and examining the impact of a particular measure. They also provide indicators into how these measures can be improved. For these reason, inputs and processes are examined alongside the intended impact on students.

At the end of an evaluation, we expect to learn something about the quality of inputs and processes, in addition to the quality of the results and impact. However, quality is a relative term that is dependent on the evaluation criteria we are using to measure it. Therefore, these guidelines provide a broad scope for expectations, standards, and criteria. We have defined what quality means in this evaluation based on a sound theoretical background and scientifically evaluated evidence, and have made an effort to provide answers to questions...

- ... On input: What do we expect from good handouts and experimentation materials? What constitutes good teacher training on conducting experiments?
- ... On process: What are good science lessons? How do experiments in the classroom best serve the learning process?
- ... and finally, on the product or result: What does scientific competency mean at a particular age level? What do we expect students do be able to do following their interaction with experimentation? How should their motivation and willingness to engage in science change?

Due to the costs involved, evaluations should be conducted efficiently and resource-consciously. Therefore, the instruments developed for this set of guidelines focus on a smaller number of indicators that are nonetheless essential and relevant for the various quality evaluation sections.

For all education stakeholders taking on the evaluation process, we hope you will enjoy a successful, purposeful, and reflective process in pursuit of encouraging children and adolescents to expand their scientific abilities, to have fun and feel confident in their scientific forays, and to become responsible stewards of our world.

Complete list of sources referenced

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

Anders, Y., Hardy, I., Pauen, S., Ramseger, J., Sodian, B. & Steffensky, M. (2013). Wissenschaftliche Untersuchungen zur Arbeit der Stiftung „Haus der kleinen Forscher“.

Arnold, J. (2015). Die Wirksamkeit von Lernunterstützungen beim forschenden Lernen: eine Interventionsstudie zur Förderung des wissenschaftlichen Denkens in der gymnasialen Oberstufe. Logos Verlag.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. Zeitschrift für Erziehungswissenschaft, 9(4), 469–520.

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 comparison of the relative effectiveness of guided inquiry and verification laboratory instruction. Science Education, 94, 577–616.

Böttcher-Graf, P. (2016). Selbsteinschätzung von Grundschullehrkräften/-studierenden in Bezug auf naturwissenschaftlichen Unterricht. Technische Universität München: Unpublished master's thesis.

Brunner, M., Stanat, P. & Pant, A.H. (2014). Diagnostik und Evaluation. In T. Seidel & A. Krapp (Publisher), Pädagogische Psychologie, 481–515. Weinheim, Basel: Beltz.

Bybee R. (2015) Scientific Literacy. In R. Gunstone (Ed.), Encyclopedia of Science Education, 944–947. Dordrecht: Springer.

Bybee, R. W., & Fuchs, B. (2006). Preparing the 21st century workforce: A new reform in science and technology education. Journal of Research in Science Teaching, 43(4), 349–352.

Cohen, L., Manion, L. & Morrison, K. (2018). Research Methods in Education. London/New York: Routledge.

Corrigan, D., Dillon, J. & Gunstone, R. (Eds.) (2011). The Professional Knowledge Base of Science Teaching. Dordrecht: Springer.

Darling-Hammond, L., Hyler, M. E., Gardner, M. (2017). Effective Teacher Professional Development. Palo Alto, CA: Learning Policy Institute.

DeGEval – Gesellschaft für Evaluation (2016). Standards für Evaluation. Mainz-Kastel: Zeidler.

DIN EN ISO 9241-11 (2016). Ergonomie der Mensch-System-Interaktion – Teil 11: Gebrauchstauglichkeit: Begriffe und Konzepte (ISO/DIS 9241-11:2015); Deutsche und Englische Fassung prEN ISO 9241-11:2015. 11. Deutsches Institut für Normung. Berlin: Beuth.

Dochy, F., Segers, M., Van den Bossche, P. & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. Learning and instruction, 13(5), 533–568.

Döring, N., & Bortz, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin: Springer.

Ehm, J.- H. (2012). Akademisches Selbstkonzept im Grundschulalter. Entwicklungsanalyse dimensionaler Vergleiche und Exploration differenzieller Unterschiede (Dissertation). Frankfurt am Main.

Ewerhardy, A. (2010). Zusammenhänge zwischen Verständnisorientierung von naturwissenschaftsbezogenem Sachunterricht und Fortschritten im Verständnis naturwissenschaftlicher Konzepte bei Lernenden der Grundschule (Dissertation). Westfälische Wilhelms-Universität Münster, Münster.

Förtsch, S., Förtsch, C., von Kotzebue, L. & Neuhaus, B.J. (2018). Effects of teachers' professional knowledge and their use of three-dimensional physical models in biology lessons on students' achievement. Education Sciences, 8(3), 118. doi: 10.2290/educsci8030118

Förtsch, C., Werner, S., Dorfner, T., Kotzebue, L. von & Neuhaus, B. J. (2018). Analysebogen Biologie 2: Kognitive Aktivierung. In C. M. Schlegel (Ed.), Schulpraktika begleiten. Stuttgart: Raabe.

Fthenakis, W. E., & Bundesministerium für Familie Deutschland. (2003). Auf den Anfang kommt es an!: Perspektiven zur Weiterentwicklung des Systems der Tageseinrichtungen für Kinder in Deutschland. G. Peitz (Ed.). Beltz.

Gichoya, D. (2005). Factors affecting the successful implementation of ICT projects in government. Electronic Journal of e-Government, 3(4), 175–184.

Guay, F., Marsh, H. W. & Boivin, M. (2003). Academic selfconcept and academic achievement: Developmental perspectives on their causal ordering. Journal of Educational Psychology, 95, 124-136.

Haslbeck, H. (2019). Die Variablenkontrollstrategie in der Grundschule. Technische Universität München: Dissertation.

Hazelkorn, E., Ryan, C., Beernaert, Y., Constantinou, C. P., Deca, L., Grangeat, M., Karikor-pi, M., Lazoudis, A., Casulleras, R. P. & Welzel-Breuer, M. (2015). Science education for responsible citizenship: Report to the European Commission of the expert group on science education. Vol. 26893. Luxembourg: Publications Office. http://dx.doi.org/10.2777/12626

Jüttner, M., Boone, W., Park, S. & Neuhaus, B. J. (2013). Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK). Educational Assessment, Evaluation and Accountability, 25(1), 45–67.

Kauertz, A., Kleickmann, T., Ewerhardy,E., Fricke, K., Lange, K., Ohle, A., Pollmeier, K., Tröbst, S., Walper, L., Fischer, H. & Möller, K. (2011). Dokumentation der Erhebungsinstrumente im Projekt PLUS. https://duepublico2.uni-de.de/servlets/MCRFileNodeServlet/duepublico_derivate_00036697/Dokumentation_der_Erhebungsinstrumente_im_Projekt_PLUS_2013_final2.pdf

Klingebiel, F. & Klieme, E. (2016). Teacher qualifications and professional knowledge. In S. Kuger, E. Klieme, N. Jude & D. Kaplan (Eds.), Assessing contexts of learning: An international perspective, Methodology of educational measurement and assessment, 447–468. Berlin: Springer.

Kobarg, M., & Seidel, T. (2007). Prozessorientierte Lernbegleitung – Videoanalysen im Physikunterricht der Sekundarstufe I. Unterrichtswissenschaft, 35(2), 148–168.

Kotzebue, L. von, Müller, L., Haslbeck, H., Neuhaus, B.J. & Lankes, E.M. (2020). Cognitive activation in experimental situations in kindergarten and primary school. International Journal of Research in Education and Science (IJRES), 6(2), 284–298.

Kotzebue, L. von & Nerdel, C. (2012). Professionswissen von Biologielehrkräften zum Umgang mit Diagrammen. Zeitschrift für Didaktik der Naturwissenschaften, 18, 181–200.

Krapp, A. (1992). Interesse, Lernen und Leistung. Neue Forschungsansätze in der Pädagogischen Psychologie. Zeitschrift für Pädagogik, 38(5), 747–770.

Krapp, A. & Prenzel, M. (2011). Research on interest in Science: Theories, methods, and findings. International Journal of Science Education, 33(1), 27–50.

Kuhn, N., Lankes, E.M. & Steffensky, M. (2012). Vorstellungen von pädagogischen Fachkräften zum Lernen von Naturwissenschaften. In H. Giest, E. Heran-Dörr & C. Archie (Publisher), Lernen und Lehren im Sachunterricht – Zum Verhältnis von Konstruktion und Instruktion, 183–190. Bad Heilbrunn: Klinkhardt.

Labudde, P. & Möller, K. (2012). Stichwort: Naturwissenschaftlicher Unterricht. Zeitschrift für Erziehungswissenschaften, 15, 11–36.

Lengyel, D. (2017): Alltagsintegrierte Sprachbildung im Elementarbereich. In: M. Becker-Mrotzek & H.-J. Roth (Publisher), Sprachliche Bildung – Grundlagen und Handlungsfelder, 273–286. Münster: Waxmann.

Lipowsky, F. (2010). Lernen im Beruf – Empirische Befunde zur Wirksamkeit von Lehrerfortbildung. In Müller, F., Eichenberger, A., Lüders, M. & Mayr, J. (Publisher), Lehrerinnen und Lehrer lernen. Konzepte und Befunde zur Lehrerfortbildung, 1, 51–72. Münster: Waxmann.

Lotz, M., Lipowsky, F. & Faust, G. (2013). Technischer Bericht zu den PERLE-Videostudien. In F. Lipowsky & G. Faust (Eds.), Materialien zur Bildungsforschung, Bd. 23/3. Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE) – Teil 3. Frankfurt am Main: GFPF.

Lüftenegger, M., Schober, B. & Spiel, C. (2019). Evaluation und Qualitätssicherung. In D. Urhahne, M. Dresel &

F. Fischer (Publisher), Psychologie für den Lehrberuf, 517–532. Berlin: Springer.

Marsh, H. W. & Craven, R. G. (2006). Reciprocal effects of selfconcept and performance from a multidimensional perspective. Beyond seductive pleasure and unidimensional perspectives. Perspectives on Psychological Science, 1, 133–163. doi: 10.1111/j.1745-6916.2006.00010.x

Mayring, P. (2015). Qualitative Inhaltsanalyse: Grundlagen und Techniken. Beltz.

Milberg, J. (Publisher) (2009). Förderung des Nachwuchses in Technik und Naturwissenschaft. Beiträge zu den zentralen Handlungsfeldern. Berlin: Springer.

Morrison, J.A. & Lederman, N. G. (2003). Science teachers' diagnosis and understanding of students' preconceptions. Science Education, 87(6), 849–867.

Müller, K., Prenzel, M., Seidel, T., Schiepe-Tiska, A., & Kjaernsli, M. (2016). Science teaching and learning in schools – Theoretical and empirical foundations for investigating classroom level processes. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), Assessing contexts of learning. An international perspective, 423–446. New York: Springer.

Nehring, A. (2014). Wissenschaftliche Denk-und Arbeitsweisen im Fach Chemie: Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung. Berlin: Logos.

Nielsen, J. (2000). Designing Web Usability: The Practice of Simplicity. Indianapolis: New Riders Publishing.

Osborne, J., Simon, S. & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. International Journal of Science Education, 25, 1049–1079.

Osborne, J. (2014). Scientific practices and inquiry in the science classroom. In N.G. Lederman & S.K. Abell (Eds.), Handbook of Research on Science Education, Bd. 2, 593–613. Abingdon: Routledge.

Patton, M. Q. (2008). Utilization-focused evaluation: The new century text. Thousand Oaks: SAGE.

Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: the German framework of Three Basic Dimensions. ZDM, 50 (3), 407–426.

Riese, J. & Reinhold, P. (2010). Empirische Erkenntnisse zur Struktur professioneller Handlungskompetenz von angehenden Physiklehrkräften. Zeitschrift für Didaktik der Naturwissenschaften, 16, 167–187.

Saffran, A. (2016). Elementary School Children's Judgments of Covariation Data: Development and Influences of Task Characteristics (Dissertation). Ludwig-Maximilians-Universität München.

Schiepe-Tiska, A., Roczen, N., Müller, K., Prenzel, M. & Osborne, J. (2016). Science-related outcomes: Attitudes, motivation, value beliefs, strategies. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), Assessing contexts of learning. An international perspective, 301–329. New York: Springer.

Schwichow, M., Christoph, S., Boone, W. J. & Härtig, H. (2016). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. International Journal of Science Education, 38(2), 216–237.

Seidel, T., Prenzel, M. & Kobarg, M. (Publisher) (2005). How to run a video study. Münster: Waxmann.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. Educational researcher, 15(2), 4–14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. Harvard educational review, 57(1), 1–23.

Smith, C. & Gillespie, M. (2007). Research on professional development and teacher change: Implications for adult basic education. Review of adult learning and literacy, 7(7), 205–244.

Steffensky, M., & Neuhaus, B. J. (2018). Unterrichtsqualität im naturwissenschaftlichen Unterricht. In D. Krüger, I. Parchmann, & H. Schecker (Publisher), Theorien in der naturwissenschaftsdidaktischen Forschung, 299–313. Berlin: Springer Spektrum.

Stergioulas, L. et al. (2014). Evaluating e-learning platforms for Schools: Use and usability, user acceptance, an impact on learning. IEEE 14th International Conference on Advanced Learning Technologies.

Stufflebeam, D.L. & Shinkfield, A.J. (2007). Evaluation theory, models and applications. San Francisco, CA: Wiley.

Tytler, R., & Osborne, J. (2012). Student attitudes and aspirations towards science. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), Second international handbook of science education, Bd. 1, 597–625. Dordrecht: Springer Netherlands.

Van Veen, Klaas/ Zwart, Rosanne/Meirink, Jacobiene (2012). What makes teacher professional development effective? A literature review. In Mary Kooy/ Klaas van Veen (Eds.), Teacher learning that matters, 3–21. London: Taylor & Francis.

Vehmeyer, J. K. (2009). Kognitiv anregende Verhaltensweisen von Lehrkräften im naturwissenschaftlichen Sachunterricht – Konzeptualisierung und Erfassung (Dissertation). Westfälische Wilhelms-Universität Münster, Münster.

Venkatesh, V. & Morris, M. G., Davis, G. B. & Davis, F. (2003). User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3), 425–478.

Vorholzer, A. (2016). Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? Eine empirische Untersuchung der Wirkung eines expliziten und eines impliziten Instruktionsansatzes, Bd.197. Berlin: Logos.

Wellnitz, N., & Mayer, J. (2013). Erkenntnismethoden in der Biologie–Entwicklung und Evaluation eines Kompetenzmodells. Zeitschrift für Didaktik der Naturwissenschaften, 19, 335–345.