Progress in Science Education
2024, Vol. 7, No. 1, page 56-77
ISSN 2405-6057
DOI.10.25321/prise.2024.1428

Research-Based Report of Practice

Development and evaluation of a GUM-based teaching-learning sequence on measurement uncertainty for upper secondary

Rupert Bärenthaler-Pachner[1], Clemens Nagel[1]

Received: March 2023 / Accepted: December 2023

Structured Abstract

Background: References to the importance of measurement uncertainty (MU) can be found in numerous national curricula. However, the number of materials for the introduction of MU in science teaching and science education research on MU in general, can genuinely be considered limited. The development of a teaching-learning sequence (TLS) for high school students (upper secondary), which is based on standardised ISO methods on how to deal with MU has not been carried out so far. Introducing MU to teachers among development and evaluation of curricula is therefore the current stage of research in this particular area.
Purpose: The present project involved the development and evaluation of a TLS for upper secondary that is based on the Guide to Expression of Uncertainty (GUM) to provide materials that can be directly implemented in the science classroom. A final product that facilitates deeper understanding of MU and its application in a scientific context by learners are the main objectives.
Design/Methods and setting: Using a design-based research approach, the project involved various research and development cycles that consisted of expert interviews, educational reconstruction, probing acceptances in addition to collection and analysis of data gathered in three different Physics classes in Austria (32 learners participated). A variety of different instruments of empirical education research (e.g. analysis of videos, interviews, quantitative data, etc.) has been used in this project.
Results and Conclusion: The central design principle was 'trustworthiness of experiments and data' following GUM recommendations of Type-A and Type-B determination. Students first build a measuring instrument by themselves to introduce Type-B uncertainties. Then - using the same instrument for their measurements -they acquire a measurement series for a subsequent introduction of Type-A uncertainties. The TLS provides a complete introduction using ISO standardised methods on how to deal with MU. The evaluation was successful, key concepts were accepted by students and learning objectives were achieved. The concept of 'trustworthiness' is currently under further investigation at the University of Vienna during a study that builds on the findings of this projects.
Keywords: *Measurement uncertainties, curriculum development, design-based research, GUM, secondary education, Nature of Science*

[1]University of Vienna
✉ baerenthaler-pachner@icloud.com, clemens.nagel@univie.ac.at

© The author(s), 2023
Published by CERN under the Creative Common Attribution 4.0 Licence (CC BY NC SA 4.0)

1 Introduction

1.1 Relevance of measurement uncertainty in (phisics and science) education

The significance of a concept for scientific reasoning and understanding must be assessed to evaluate its implementation in the science classroom. Measurement Uncertainty (MU) is a fundamental part of experimental scientific practices and the generation of knowledge, which makes it essential for doing science in general. Such standardised practice regarding MU is as of now also demanded by an ISO standard, which is based on the Guide to Expression of Uncertainty (GUM). Experimental methods are not only part of academic research, but also account for a central part of physics teaching (Harlen, 1999; Heinicke, 2014; Tesch & Duit, 2002). Harlen (1999) concludes that learners need to "do science" (p. 23). in order to understand it and highlights the role of experiments in physics education. Heinicke (2012, p. 4) gets to the heart of this matter by stating that teaching the concept of MU correspond with showing the adequate limitation of scientific knowledge on both a tertiary and secondary education level. The latter is explicitly put into the context of contributing to general education purposes and establishing responsible citizenship. This aligns with the potential of MU and deeper understanding of Nature of Science (NOS). Buffler et al. (2009) elaborate on the relation between knowledge about the "Nature of Scientific Measurement" and about NOS in general. Heinicke (2012, p. 12) adds that aspects of MU are valuable information rather than shortcomings or flaws of an experimental result, which are decisive for its interpretation and analysis. This circumstance is highlighted by the fact that the interpretation of results is a key element of every scientific protocol. Scientific comparison or distinction of two measurement values, for instance, would not be possible without consideration of MU.
A development that has become perhaps more visible recently is public mistrust of science (Millstone & Zwanenberg, 2000). It can be argued that education systems all over the world fail to teach about basic scientific limitations and methodology. This can be considered one factor that leads to some form of disappointment and ultimately critical attitudes towards science in general for a significant number of the public during the COVID-19 pandemic (Rowland et al., 2022; Van der Bles et al., 2020; Van der Linden & Löfstedt, 2019). Under these circumstances, the implementation of MU in the science classroom could have the potential to prevent such attitudes by contributing to a much more veritable understanding of NOS. It could be argued that questions regarding the existence and determination of the true value of a quantity are part of a rather philosophical discussion about experiments. However, if the different arguments above are concerned, the determination of a quantity with its corresponding uncertainty is a key foundation of (the philosophy of) physics and science, which makes it a key concept for science education. MU is substantially linked with basic principles of acquisition of knowledge and the theory of falsification (Popper, 1934). Buffler el al. (2009) found such positive influence of learners' understanding of MU on the conceptualization of NOS.

1.2 Research gap in this specific field

Despite a considerable number of references to MU in various national curricula (e.g. Federal Ministry of Education, Science and Research, 2017; Department for Education of the UK, 2015), a study conducted by Nagel et al. (2021) displayed that teachers in the Austrian education system rarely implement MU in their teaching. Stated reasons were insufficient competences regarding MU by the teachers themselves and a lack of available teaching materials or corresponding chapters in schoolbooks. Nonetheless, the study also showed that those teachers consider MU to be an important topic and a meaningful part of general education. Corresponding findings are provided by Möhrke (2020) based on a teacher survey in Germany (State of Baden-Wüttenberg). This circumstance is also reflected by only a small number of references to existing evidence-based materials on MU in recent publications (cf. Unterrichtskonzeptionen für den Physikunterricht" Schecker et al., 2021). Research that has been conducted so far, will be elaborated on in section 2.1.1. as some materials on MU have been developed rather recently. Such works have focused on isolated examples for application of MU (Hellwig et al., 2017) or exclusively on the development of a normative comprehension model that focuses on subject matter on MU relevant for secondary education (Hellwig, 2012). However, no specified teaching-learning sequence (TLS) or curriculum for upper secondary based on GUM has been developed and evaluated up to this point (Schecker et al., 2021).
What can be observed in general is that recent projects and research on MU and education indicate a change in trend within science education research. This is represented by research focussing on the development of conceptual models on MU and approaches to the fundamental introduction of its relevance within science education research (Heinicke, 2012; Heinicke et al., 2010; Heinicke & Holz, 2018; Heinicke & Holz, 2020; Hellwig, 2012; Hellwig & Heinicke, 2020, Priemer & Hellwig, 2018). Additional worked indicating the described shift in focus relates to practical courses at university level (Buffler & Lubben, 2008; Petts et al., 2021) or the field of students' beliefs and the relationship between knowledge about MU and different competences of learners (Kok & Primer, 2023; Kok et al., 2019; Heinicke & Holz, 2020; Masnick & Morris, 2008; Nedden & Priemer, 2020). However, Nagel et al. (2021) concluded based on obtained data, that additional (didactic) research and the development of evidence-based materials is necessary for a long-term implementation of MU by teachers. So, a broad research and development project at University of Vienna started to systematically develop and evaluate curricula for the introduction of MU in lower and higher secondary (Nagel, 2023). One TLS for lower secondary (age 10-14) based on GUM was developed by Loidl (2021) during a Master's project at the University of Vienna. An additional one for upper secondary was developed by Bärenthaler-Pachner (2022) and is

the matter of this paper.
Preceding elaborations outline the significance of MU in science education and that its broad implementation has influence on more general aspects of scientific understanding that cannot be anticipated or evaluated as of now. It can be argued for a change in attitude towards MU in the field of physics teaching as brought to the point by Heinicke and Holz (2019) as they refer to MU as an unwanted visitor in class figuratively standing between scientists and their desired true value, which should be given a different and scientifically more valuable role.

2 Research background

2.1 Theoretical key concepts

There is a close and unfortunately sometimes ambiguous connection between the concepts of uncertainty and error. Heinicke (2012, p. 9) casts light upon historical reasons and linguistics issues for this circumstance. Terms such as error or propagation of error derive from the fundamental model of Carl Friedrich Gauß (*Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*) and have been used without clear consistency at university level. It should be noted that already Keppler used the term *incertitudo* to show that the term is a semantically and conceptionally not a merely recent invention (Heinicke, 2012, p. 92). It may be suggested that no clear semantic distinction between error and uncertainty has at least the potential of confusion for learners at secondary level. That is why science education researchers from this particular domain (Buffler & Lubben, 2008; Heinicke & Holz, 2019; Loidl, 2021; Nagel, 2021; Petts et al., 2021; Priemer, 2022; Schecker et al., 2018, etc.) are in favour of the terminology suggested by GUM (BIPM, 2008), which corresponds with the ISO/IEC Guide 98-3:2008-09 norm. Basic principles and concepts established by GUM are seen as the appropriate basis for the development of concepts and materials for science education on tertiary and secondary level, despite some criticism (Buffler et al., 2009; Heinicke, 2012; Nagel, 2021). Nagel (2021) and Heinicke (2012, p. 247) highlight not only the didactic, but also economic value of GUM for international collaboration in research and business by comparing it to the International System of Units (SI) in that regard.
The essential distinction between errors and uncertainties is that errors have to be avoided, while uncertainties are unavoidable and essential to estimate the probability of the existence of the true value within the range of uncertainties. If an error produces a measurable difference between a measurand and a reference standard, it is called systematic error and has to be corrected (Nagel, 2021, p. 8). GUM defines "uncertainty arising from random effects and from imperfect correction of the result for systematic effects" and that the "exact value of the error in the mean arising from these effects cannot be known" (p. 5). It is also mentioned explicitly that the "uncertainty of the result of a measurement should not be confused with the remaining unknown error" (BIPM, 2008, p. 6).
It can be summarised that the main difference between determinable errors und uncertainty is that MU in contrast to such errors cannot be eliminated. (Nagel, 2021, p. 8). GUM (BIPM, 2008, p. 5) therefore suggest the term "random effects" instead of random error to avoid confusion.
Sources for MU are provided in (BIPM, 2008) and include "inexact values of measurement standards and reference materials, finite instrument resolution or discrimination threshold" and "approximations and assumptions incorporated in the measurement method and procedure" (p. 6) amongst other examples. MU is closely linked to the measurand, which is defined as the "quantity to be measured" (BIPM, 2008, p. 49). It related to the concept of true value. According to GUM a distinction between the measurand and the "true value of the measurand (or quantity)" (BIPM, 2008, p. 4) is not necessary. Nonetheless, the concept of true value will be presented in further detail in this section due to its relevance for the TLS based on didactic assumptions (section 4.1). Before elaborating on issues regarding the true value, the question needs to be answered how MU can be determined.
There are two different types of evaluating MU according to GUM. Type A evaluation (of uncertainty) is based on the "statistical analysis of a series of observations" (BIPM, 2008, p. 3) which involves the application of a suitable model (e.g. Poisson distribution, Gaussian distribution, etc.) for a specific series of measurements. If the sample size n of this series is large enough ($n > 10$), the type A evaluation of uncertainty (TAU) is equivalent to the standard deviation of the (arithmetic) mean u_x, and defined as shown in (1).

$$u_x = \frac{s_x}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n \cdot (n-1)}} \tag{1}$$

Standard deviation of the mean, also referred to as standard error[1] (Nagel, 2017, p. 29), is the appropriate statistical quantity to describe the dispersion of the mean since the dispersion of an individual data value to the mean (i.e., standard deviation) is of no relevance for future predictions. A future mean will appear in an interval of $\pm\ u_x$ around the mean with a probability of 68,3 % and likewise the true value lies within the same area and the same probability according to this model (Nagel, 2017, p. 30).
Type B evaluation (of uncertainty) is defined as a "method of evaluation of uncertainty by means other than the statistical analysis of series of observations" (BIPM, 2008, p. 3) and can be provided along with one single measurement. Any gauge that includes a scale, accounts for type B evaluation of uncertainty (TBU) that is composed of three main subtypes (i.e., uncertainty of calibration, linearity and reading) among other factors that must be considered. Its operating mechanism, material properties and method of production can cause several undetermined deviations between measurements. Consequently, TBU hast to be provided by the manufacturer directly on a gauge or included in the manual (e.g. digital multimeters). The value for TBU frequently corresponds with the reading uncertainty, which is either the last significant digit displayed of a digital reading or the resolution of an analogue scale. TBU can be seen as the accuracy of a measurement, whereas TAU as its precision. In general, accuracy is of greater value when it comes to providing information about measurements, unless the TAU is larger than the TBU. To prevent providing an unattainable MU with the final result of a measurement, the larger uncertainty is relevant for each specific experimental setting. The result therefore consists of the arithmetic mean of a series of observations and the appropriate MU.
A final key concept is the true value of a quantity, which cannot be determined according to basic principles of empirical science (Nagel, 2021, p. 9; BIPM, 2008, p. 32). GUM addresses this aspect by classifying the word 'true' as redundant and simply referring to the value of a quantity or measurand (BIPM, 2008, p. 50). However, this does not rule out its existence. On the contrary, statistical models show why the true value has to exist and why the mean, determined on the basis of a set of data, is its best approximation. Likewise, Fig. 1 displays that standard deviation of the mean is the appropriate statistical quantity in contrast to standard deviation within the model. The beauty of this model lies in its characteristics that the deviation of the mean will be zero for an infinite number n of measurements (Fig. 1), which would lead to the value of the mean and the true value being identical.

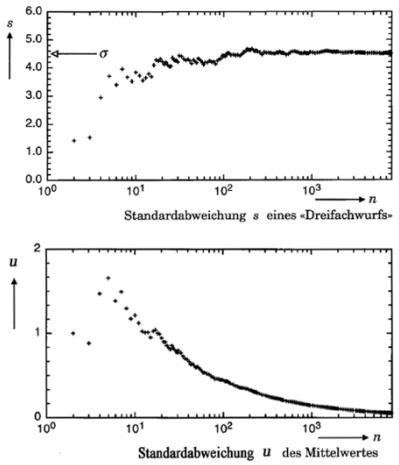

Fig. 1. Comparison of standard deviation (top diagram) and standard deviation of the mean (bottom diagram) for a dice experiment (Gränicher, 1996, p. 27)

A detailed summary of theoretical concepts that are relevant for the topic of MU is provided explicitly for teachers by Nagel (2021), which also elaborates on further aspects of TBU, differences between precision and accuracy, measurand, types of errors, etc.

2.2 Summary of relevant literature on MU

2.2.1 Current state of research and curricular legitimations

As outlined, the amount of research in the field of MU and science education has been rather limited until recently. Heinicke and Holz (2019) describe MU as an unwelcome visitor, which can be seen as linked to teachers being in favor of less sensitive experiments, which do not involve large MU (Höttecke, 2013). Recent research and the development of teaching concepts has been carried out by Buffler and Lubben (2008), Glomski and Priemer (2010), Heinicke et al. (2010), Heinicke and Holz (2018), Heinicke and Holz (2019), Lippmann (2003), Loidl (2021) and Petts et al. (2021).
As not only content knowledge about MU, but also specific didactic competences have been regarded as decisive for

[1] This represents another example of terminology that can lead to ambiguity and confusion between uncertainty and error.

the implementation of this topic in science education (Buffler et al., 2001; Heinicke & Riess 2009), the development of corresponding materials and in-service teacher training have significant potential. However, there are no known curricula, which are specifically based on GUM and designed for upper secondary learners so far. The limited number of materials and first concepts for the introduction of MU in upper secondary that can be identified were developed by Boczianowski & Kok (2020), Kok & Priemer (2023), Neumann (2021) and Wagner et al. (2021). Justifications for fundamental understanding of MU can be found in various legal frameworks for different education systems. The Austrian national curriculum for instance explicitly refers relevant aspects such as observation, description, analysis and protocolling experiments, understanding of methods and practices specific to physics and teaching about scientific models in physics (Federal Ministry of Education, Science and Research, 2017). National curricula for upper secondary in Switzerland also refer to the importance of topics and concepts that can all be seen as related to methods and practices in physics, which undoubtedly require knowledge about MU. (EDK, 1994, p. 108; Department for Education - Canton St. Gallen, 2019, p. 65). The curriculum for the canton of Bern also clearly lists MU as a key concept for the beginning of physics teaching and providing measurement values in an appropriate way (Department for Education - Canton Bern, 2017, p. 197). As far es German curricula are concerned, relevant documents from various federal states widely include references to experimental competences and knowledge about scientific practices as well (Ministry for Education, Youth and Sport – Brandenburg, 2021; Ministry for Education - Rheinland Pfalz, 2022). The curriculum for the state of Bavaria, which is planned to become applicable in 2025/26, accounts for detailed and explicit references to MU. MU is linked to comparing and evaluating scientific results, the significance or meaningfulness of data and the process of acquiring knowledge in physics and science in general (Federal Institution for Education Standards and Research in Education - Munich, 2023, p. 2-4). Regarding an English-speaking context, the national curricula for GCE AS and A level in physics too state that learners should be able to "evaluate results and draw conclusions with reference to measurement uncertainties and errors" (Department for Education of the UK, 2014, p. 19) and the "limitation of physical measurement" is listed as required knowledge and understanding (Department for Education of the UK, 2014, p. 12). Despite considerable differences between federal states within the United States, the National Research Council, which is a part of the National Academy of Sciences, publishes the "Framework for K-12 Science Education" that is seen as significant influence on science education standards and curriculum development in the US. MU is referred to in the context of what it "means to learn science" and "understanding the nature and development of scientific knowledge" (National Academies of Sciences, 2012, p. 251). "Ways of dealing with uncertainties and agreed-on levels of certainty" are considered crucial elements of scientific knowledge and skills (National Academies of Sciences, 2012, p. 251). It can be summarised that the importance of MU for deeper scientific understanding of basic principles is part of numerous curricula on an international scale. More recently updated documents tend to go even further into detail when acknowledging the concept of MU with regards to competences and knowledge that can be linked to NOS. It can be argued that such deeper understanding is without exaggeration fundamental to education in physics and science. Additionally, official documents of the OECD also explicitly refer to the importance of understanding MU in the context of scientific reasoning (Strategic Visioning Expert Group, 2020) to name another official document within a greater context. This document on a strategic vision for science education names the concept of uncertainty as a "central feature of most scientific issues" relevant to all generations (Strategic Visioning Expert Group, 2020, p. 9). Ways of expressing uncertainty as a "key element of scientific culture" are considered vital to respond both individually and as a society to a variety of challenges which, for instance, could be "posed by global health emergencies, such as a pandemic" (Strategic Visioning Expert Group, 2020, p. 9). Furthermore, probabilistic assertions, which involve the determination of MU, are seen as an opportunity to teach young people about science's strengths rather than weaknesses, by facilitating a change of position with respect to new evidence. This corresponds with a statement by recent Austrian Nobel laureate Prof. Anton Zeilinger, describing the greatness of science lying within the circumstance that a whole scientific community can change its opinion if there is experimental evidence disproving models and theories that sometimes even have been established for centuries (Anjobi Videoarchiv, 2023).

To conclude, from a traditional point of view, the role of the experiment in science education has further evolved. Undeniable central in physics teaching, its role changed from the art of demonstration (Wiesner et al., 2011, p. 113) towards a learning process that puts learners' participation and action in the centre of experimental settings. In that way learners should be given the opportunity to explore and work with new concepts, such as MU, to facilitate the development of competences. Another aspect mentioned in the Austrian national curriculum is the relevance of physics for the economy. This can be related to MU as outlined before. The consensus in science education research, that the approach provided by GUM is the most suitable in science education has already been mentioned before. Nagel (2021) puts emphasis on this by stating that the distinction between error and uncertainty provided with this ISO-norm and its inherent use of consistently precise language has to become ubiquitous practice in science and teaching. The relevance of MU for better understanding of NOS has been elaborated on in 1.1. and this section. References can be found in multiple official documents and the current state of research also indicates benefits of learning about MU for learners of science. At this point it can only be assumed how significant the implementation of MU as a fundamental concept affects future generation´s views on our world, our relationship to physics and the public perception of science in general.

2.2.2 Research on students' beliefs

Eine Data on (mis)conceptions about MU is likely to be linked to conceptions about NOS in general, which is why research in this area is also relevant. Schecker et al. (2018, p. 279) describe that learners believe that the laws of physics exist in the world and doing science means trying to find them. Heinicke (2012, pp. 477- 478) also refers to the attitude of students at school, university level and even teachers that science is objective, and that scientific reasoning does not involve human conclusions, creativity, or the consideration of the theoretical framework of a model. This shows that the model-based nature of science does not seems to be part of the general conceptions for the majority of learners. Anecdotic evidence of similar disbelief can be seen by the reaction of university students in one of Richard Feynman statements during one of his Messenger Lectures on "The Character of Physical Law" in 1964. The audience laughed when Feynman elaborated on the very first step of building a scientific model with "[f]irst, we guess it, [audience laughs] no don't laugh that's the truth." (Burton, 2015). This example shows that even for university students of physics back then, this aspect of NOS seemed far from obvious and recent data shows the same patterns.
As far as specific conceptions about MU are concerned, Schecker et al. (2018, p. 279) argue that beliefs about experimenting include views on measurement and its role in an experimental setting. This involves the idea that a measured value and especially the process of measurement is straightforward and one single measurement provides, if executed properly, a true and accurate result. Buffler et al. (2001) describe this "point reasoning" as the view that each single measurement is seen as one possible result for the true value of a quantity. Heinicke (2012, p. 473) points at the relationship between such ideas and the dominating conception of feasible determination of the true value. Furthermore, Schecker et al. (2018, p. 279) argue that the credibility or trustworthiness of one measurement is highly overestimated. The usage of the term error (of measurement), or (measuring) error or similar wordings are estimated to reinforce this conception, which is why Schecker et al. (2018, p. 279) also argue for the term uncertainty from a didactic point of view. A study by Heinicke (2012, p. 49) showed that MU was considered unnecessary information of an experimental result for most university students. Evangelinos et al. (2002) observed in another study that university students would only provide MU with results, which were considered not to be trustworthy. Furthermore, Ryder and Leach (1999) detected the idea that the quantity of measurements is associated with its quality at university level. Garret et al. (2000) found similar lack of distinction between the concepts of precision and accuracy for chemistry students.
Loidl (2021, p. 59) concludes based on her data, that learners at lower secondary level associate MU with the experimenter instead of the experimental setting. Although the number of research on students' beliefs is again limited, main characteristics indicated by significant research can be summarised as follows:

- The true value of a quantity can be determined (if the scientific setting and execution are optimised)
- A single accurate measurement provides the true value
- Errors/uncertainties can be eliminated
- Error and uncertainty describe the same phenomenon
- MU do not have to be provided with the result of a measurement (unless it can be considered 'flawed')
- MU is caused by human factors and are not part of NOS
- Trustworthiness of a result is linked to the person or group of people who conducted the experiment.

The issue of no clear distinction between the concepts error and uncertainty lies at the heart of most listed views and is likely to stem from the outlined inconsistent use of different terminology by educators and scientists. This lack of distinction can be regarded as closely connected to conceptional beliefs about the true value. These are likely to be caused by associations with the words error and true value, while neglecting or at least overlooking aspects of MU. However, due the low frequency of the term uncertainty (germ. *Messunsicherheit*) in semantic contexts of learners at secondary level and everyday language in general[2], it can be argued that learners have no stable pre-existing concepts for MU at least for German and English-speaking contexts. That is also why the approach of a conceptual change (Duit & Treagust, 2003; Schecker et al., 2018) cannot be considered suitable in comparison to the opportunity of establishing a new concept called MU. This also complies with the approaches suggested by Haagen-Schützenhöfer and Hopf (2020, p. 9) outlining that conceptual change "is triggered by continuous strategies rather than by cognitive conflict". The implementation of "key stimuli that activate desired knowledge elements" Haagen-Schützenhöfer and Hopf (2020, p. 9) is advised instead. Consistent use of concise language in the field of MU relates to such stimuli and is also promoted in the relevant literate (Heinicke, 2012; Nagel, 2021; Schecker et al., 2018).

[2] A frequency of 6 per million words in the spoken part of the British National Corpus for *uncertainty* is for instance provided by the University Centre for Computer Corpus Research (University of Lancaster) https://ucrel.lancs.ac.uk/bncfreq/flists.html

3 Methodology

3.1 Design based research (including the Model of Educational Reconstruction (MER)

Findings of science education research can unfortunately still not be considered of determining influence on the practice of physics teachers (Haagen-Schützenhöfer, 2015, p. 4). Duit et al. (2012, p. 15), for instance, observed that science education is considered irrelevant by a majority of teachers. It is argued by Gräsel and Parchmann (2004, p. 204) that a transfer of science education knowledge into teaching practice is more likely if evidence-based teaching concepts take school-specific circumstances into account and abide by them. This aspect of practicability leads to a conclusion by Haagen-Schützenhöfer (2015, p. 4) that the goal of science education research should not only focus on improving methods and results, but rather be motivated by the needs and issues of such practical aspects of teaching. Duit et al. (2012, p. 15) argue for the method of Design- Based Research (DBR) to develop scientifically evaluated materials that are derived from school realities.

With this understanding, DBR can be seen as a suitable method to affect both future research and teaching practices in a beneficial way. Haagen-Schützenhöfer (2015) provides a corresponding application for "design-based research as a model for curriculum development" in introductory optics. DBR can be considered of qualitative nature (Feulner et al., 2015, p. 217) with a characteristic interaction between methods, materials, media, teacher(s) and learners (Reinmann, 2018, p. 12). Despite DBR not being fully compatible with the quantitative criteria of objectivity, reliability and validity (Altrichter et al., 2018, p. 104), this does not imply that qualitative methods and tools cannot be included in DBR projects. However, the six quantitative criteria provided by Mayring (2016, pp. 144-148) can be regarded as suitable for DBR, which lead to the development and selection of most testing instruments of this project. DBR is characterised by its variable application of research- and development cycle (RC & DC), in which processes can take place on a heuristic, empirical, production and validation level (Haagen-Schützenhöfer & Hopf, 2020). At its core lies the teaching and learning environment (TLE), which can be seen as an evolved final TLS, and the aim of DBR is the development of such a final product that has scientific and practical relevance. The superordinate research question of such a research method can be seen as, whether the TLE facilities the achievement of predefined learning goals. The strength of DBR for development in education in general lies in its application within "naturalistic" contexts while combining "empirical education research with the theory driven design of leaning environments (Haagen-Schützenhöfer & Hopf, 2020, p. 2). It can be argued that this method has the potential of becoming a cornerstone of further teacher professionalisation and a standard practice for research project during teaching degrees (e.g. Bachelor or Master papers) in general.

However, there is criticism about potential limitations of DBR regarding issues of generalization and what is referred to as the paradigm of DBR (Haagen-Schützenhöfer & Hopf, 2020, p.1), which includes the approach often being "criticized for having unclear methodologies for warranting claims". This paradigm includes design principles guiding the design process and additionally also functioning as hypotheses "that are tested to refine local instruction theories about teaching and learning" (Haagen-Schützenhöfer & Hopf, 2020, p.4). Guisasola et al. (2023, p. 23) comment on this issue by stating that DBR seeks to "study the efficacy of particular TLS designs and, on the other hand, develop humble theories on classroom science teaching", which can be considered humble due to context specific analysis of teaching-learning processes. The importance of implementing different methodologies to generate general theories implies higher aspects of generalisability with a decrease of testing instruments and research- and development cycles. These humble theories are sometimes also referred to as "local theories" (Haagen-Schützenhöfer & Hopf, 2020; McKenney & Thomas, 2018) within the context of educational design research. McKenney & Reeves (2018, p. 35) point out that a "local theory is produced when limited manifestation of a certain phenomenon is studied (e.g. several iterations of one basic intervention are studied in just a few classrooms)". It is described that they can be used for the development of similar design studies and thus potentially induce further research in general. Despite outlined issues, these can also be seen as one strength of DBR, as in opposition to a traditional theory-testing setting, a DBR approach "explicitly exploits the design process as an opportunity to advance the researchers understanding of teaching, learning, and educational systems" (Edelson, 2002, p. 107). Aspects of representativeness and generalization must therefore be reflected on when conducting and presenting results and products within a DBR context (Brown, 1992, p. 173; McKenney & Thomas, 2018; Confrey, 2005, p. 147). Stake (1995) refers to interpretations of such reflections and presented data as "petite generalizations" to introduce a suitable metaphor.

DBR may not only serve as a tool for further innovation in education, but also greater public recognition of the teaching profession due to such field specific publications of practical relevance. Cobb et al. (2003) touch upon this by putting emphasis on the precondition that when conducting a DBR project "the theory must do real work" (p. 10).

Despite Haagen-Schützenhöfer (2015, p. 21) mentioning the Model of Educational Reconstruction (MER) based on Kattmann et al. (1997) as one of many components of DBR, it should be noted that conducting MER on MU played such a fundamental role for the design of the TLS, that it should be mentioned explicitly as another main method in

this section. The methodology of this particular DBR project on MU involved the following research- and development cycles (Tab. 1):

Tab. 1. Summary of research and development cycles according to DBR

cycles	**description and setting**
DC I	MER (Kattmann et al., 1997) and development of a first version of TLS
RC I	Expert interview (Bogner et al., 2009) and analysis of TLS (Dr. Clemens Nagel)
Micro-DC	Testing of experimental settings of TLS with 2 groups (NMS-Obervellach, Carinthia)
RC II	Expert interview (Bogner et al., 2009) on adapted TLS (Univ.-Prof. Dr. Martin Hopf)
DC II	Probing acceptances (Jung, 1992) with 3 learners (Bundesgymnasium 18 Klostergasse, Vienna)
DC III	Testing of TLS in classroom setting, 2 groups (Bundesrealgymnasium 14 Linzer Straße, Vienna)
RC III	Presentation of first data and TLS during expert discussion (AECC-Physics, University of Vienna)
DC IV	Final testing of TLS in classroom setting, 1 group (Bundesrealsgymnasium Lienz, Tyrol)

The development of teaching materials and an overall teaching concept for MU involved the formulation of operationalised learning goals. A number of 6 testing instruments (cf. triangulation; Mayring, 2016, pp. 144-146) were utilised during evaluation cycles of the DBR project. Those instruments were partly specifically designed or adapted in accordance with mentioned goals. These were formulated and accessed during MER (DC I), the two expert interviews (RC I & RC II) as well as the expert discussion (RC III) and are listed in section 4.1 (Tab. 2).
DC I involved analysis and summarisation of the relevant literature for both specific terminology and relevant findings on MU in science education. Core concepts, which will be elaborated on in the following section were defined and first drafts for a TLS on the basis of teaching models (Oser and Baeriswyl 2001) were developed. The concepts and materials were presented and discussed in RC I and subsequently the experimental settings were tested with two different groups to evaluate anticipated time frames and issues of practicality (Micro-DC). The Micro-DC included observation and reflection materials for the teacher who tested the materials along with additional feedback by the teacher after the lessons. At this stage of the project, adapted materials and preliminary results were again presented and discussed during an expert discussion in RC II. The following two cycles involved probing acceptances with three different learners. Developed materials were tested with those three students and additional follow-up questions were discussed individually with each of them (e.g. motivational and practical aspects, evaluation of terminology by the learners, definition of core concepts, etc.). Hypotheses that had been formulated on motivational aspects, practicality and learners' conceptions on MU were confirmed within this setting. DC II and especially DC III correspond with the validation level that is called "ecological validation of artifacts" (Haagen-Schützenhöfer & Hopf, 2020, p. 4) which refers to a DBR product being tested in teaching settings that are as authentic as possible. During DC III, the TLS was conducted with two different groups, which involved following analysis of video data and pair work video data, written assessments, analysis of observation forms for an additional teacher (passive observer) and analysis of in class questions and tasks via an online tool. Before the expert discussion in RC III that involved researchers from the AECC- Physics in Vienna, steps and results from all previous cycles, the current version of the TLS and collected data were presented to those experts. Additional questions were provided to start the expert discussion and participants additionally elaborated on further aspects (section 4). In DC IV, the same approach was employed as in DC III to collect supplementary data within a distinct environment. Conclusively, an analysis and comparison of all the gathered data was conducted to assess whether the set learning objectives could be deemed achieved within the scope of this DBR project.
To summarise, testing instruments consisted of probing acceptances, written assessments, tasks during the lessons (data collection via online tool), two different types of observation forms, post-teaching discussions with class teachers (cf. communicative validation; Mayring, 2016; Klüver, 1979), analysis of classroom video data and analysis of pair work video data (cf. categorised date analysis; Altrichter et al., 2007; qualitative content analysis; Mayring, 2000). The written assessment was designed to match single items with one specific predefined learning goal and can be considered a tool of quantitative nature compiling with the corresponding criteria. The initial phase of DBR additionally focuses on the deduction of design principles from theoretical and empirical foundations (Haagen-Schützenhöfer & Hopf, 2020). Refinement took place in the same cycles as it was the case for learning objectives, which are listed and elaborated on in the following result section as they can be seen as a decisive constituent of the final TLS.

4 Results

This section presents the two main areas of outcome of the research process, which are the main properties of the final version of the TLS and the presentation of data that has been obtained during several cycles.

4.1 Key characteristics of DBR product (TLS on MU)

The basis for the development of the teaching concept were methodology-based design principles on a general level and in consequence key ideas and operationalised learning objectives for the development of teaching materials. Design principles were based on relevant previous DBR works, expanded on the basis of research cycles and also resulted from the findings of MER on MU (Tab. 2). The principle of practicability related to the DBR requirement of impacting teaching and learning in "naturalistic context" (Barab & Squire, 2004, p. 2). Practicability responds to the reality of limitation regarding materials, (prep)-time and the general setting in class. Thus, this principle demands a minimum of available materials for experimental phases so that further use by practitioners is promoted. Haagen-Schützenhöfer and Hopf (2020, p. 2) relate to this aspect of DBR by stating that "effectiveness and easy implementation in authentic classroom settings" is strived for, which also requires involvement of practitioners from early development stages on. The development of one worksheet packet that includes tasks, exercises, essential definitions of key concepts and additional information means that the relevant content for the topic of MU is covered by one document. Such a packet was designed to be universally applicable in lower secondary due to no standard establishment of MU at the start of secondary physics education yet (Nagel et al., 2021). This manifests itself essentially in the aspect that no additional physics-specific content is imbedded in the TLS, and that no new science knowledge is established alongside MU or required to work on the packet. However, implementation is strongly advised at an early stage of physics teaching and therefore also designed for the beginning of secondary education. In that way MU can be included in experimental tasks from an early stage of physics teaching and competences extended with following experiments in class. Materials for the whole TLS include this worksheet packet, detailed lesson plannings, a didactic commentary (teacher's guide) and print templates for the required materials.

The definition of core concepts, which are *MU (Type A and B), measurement value, measurand, mean, true value* and *unit*, stem from MER and research cycles. These were derived from relevant research on characteristic concepts that has been carried out so far (e.g. Buffler & Lubben, 2008; Heinicke, 2012; Kok, 2022) during three research cycles. The concept of the true value, for instance, is also part of materials created by Buffler et al. (2005). The concept of *unit* is included in an optional additional part of the TLS that has been developed but not evaluated. The idea behind this optional lesson is to cover topics (e.g. units, orders of magnitude and the SI) that have already been implemented by teachers in their practise. This approach was taken to motivate teachers to incorporate MU into a topic they are already planning to teach.

Additionally, the concept of "trustworthiness" (germ. *Vertrauenswürdigkeit*) was adopted by Loidl (2021) and is given a more central role for the development of this particular TLS. The concept and wording of trustworthiness therefore is mirrored explicitly in learning goals, throughout the entire worksheet packet and testing instruments were signed to focus on the concept too. Trustworthiness as a tool to combine everyday language of learners with the concept of MU can be interpretated as a local theory (section 3.1) that has been applied and refined within this DBR project. The function of trustworthiness is assumed to connect terminology used by learners in daily interaction and the technical term MU. Despite obvious linguistic reasons with differences across multiple languages, it can be argued that within a German-speaking context and likely the English language as well, the introduction of *Vertrauenwürdigkeit* or *trustworthiness* embodies additional support for conceptual understanding of MU. The relation between trustworthiness and MU is therefore also elaborated on by Covitt & Anderson (2022) and the concept of trustworthiness is seen as relevant to "developing science literacy" and learners should engage with both trustworthiness and MU "separately and together" (Covitt & Anderson, 2022, p. 1175). Pols et al. (2019) likewise use the term trustworthiness during a series of experimental activities with upper secondary learners to evaluate results and the quality of their measurements. For science teaching within other linguistic contexts, science education research within that specific community on terminology is obviously required when designing or adapting teaching materials on MU. Hu and Zwickl (2018) also introduce the trustworthiness of experimental results within a survey examining students' views on validity of experiments at university level. Likewise, Fussel et al. (2022, p. 194) adopt the same items on trustworthiness for "a survey question that probes students' perspectives on the reliability of physics experimental results". This was used within the context of machine learning for automated content analysis.

The principle of including appropriate statistical analysis (e.g. calculation of mean and deviation of the mean) according to GUM was formulated due to two main reasons. Usage of mathematical tools such as GeoGebra, Excel or advanced calculator functions enable learners to calculate these values within a short period so that statistical analysis will not be the main activity of a lesson. Secondly, implementation of proper tools will avoid having to replace previous concepts with more accurate ones subsequently. The concept of the deviation of the mean and its relation to the true value is considered of such conceptual significance that it must be included to reach defined learning goals. The entirety of leaning objectives in accordance with key ideas are listed in Tab. 2 and the entire design principles are listed below:

Design principles

- The central conceptual term is **'trustworthiness'** of results and its determining factors (e.g. number of measurements, errors during the conduction of an experiment, specification of a certain gauge, etc.).
- Practicality
- Context-independent use of the TLS packet throughout upper secondary level
- Misconceptions or learners' beliefs are not explicitly addressed or presented.
- There are no significant preconceptions about MU among learners yet, so this concept can be developed without direct reference to learners' perspectives on MU[3].
- The instructional subject matter is not determined by the scientific subject, but by the needs of the learners (Haagen-Schützenhöfer, 2015, p. 26).
- The focus of the TLS is to make learners aware that the indication of MU is a fundamental component of measurements and scientific practice.
- The Fundamental scientific terms (cf. Elementaria; Kattmann et al., 1997) are MU (Type A and B), measurement value, measurand, mean and true value.
- The Fundamental concepts are the 'trustworthiness' of measurements (factors that influence trustworthiness) and the scientifically correct reporting of results using recognised formalisms.
- Statistical analysis is not the main focus of the TLS and should occupy only a small part of it. Nevertheless, the technically correct evaluation according to GUM is preferred, as this should not pose any difficulties for learners using mathematical tools (e.g. GeoGebra, calculator, Excel, etc.), and additionally, no concept is introduced that may later have to be discarded.

Tab. 2. Key ideas and operationalised leaning objectives resulting from MER and expert interviews

key ideas	learners can … (operationalised learning objectives)
The individual measurements in a series of measurements are usually not identical (Loidl, 2021, p. 90).	… describe steps during the production and use of a measuring tool that lead to TBU.
MU can be determined using method A and B.	… differentiate between TAU and TBU. … calculate the mean and standard deviation of the mean for a given set of measurements.
Results can differ in terms of their trustworthiness.	… evaluate measurements based on their trustworthiness. They can then rank different measurements based on their MU and make comparisons regarding trustworthiness.
A measurement value should always be indicated with the appropriate MU.	… correctly report measurement results (mean ± uncertainty) and distinguish between correct and incorrect indication of scientific measurement values.

[3] This is only the case within a German-speaking context since germ. *Messunsicherheit* represents a technical term from the field of experimental science only and has no additional meaning or interpretation. The German word *Unsicherheit* on the other hand does not correspond with the semantic field and associations related to *uncertainty* in English. Instead *Unsicherheit* resembles the feeling of insecurity within the spoken German language (cf. germ. *sich unsicher fühlen*) and connotations within a scientific context can be considered unlikely for the target group.

4.2 Main structure of TLS

Die The TLS is designed of two lessons and consists of two coherent sessions followed by a third optional one. The optional lesson follows the TLS in methodology and is about the topic of (SI)-units, dimensions and orders of magnitude. These specific materials were created to make the TLS more appealing for practitioners as these concepts are usually covered at the start of most physics' classes at least to some degree. In that way, teachers are enabled to teach a topic that is already implemented both in curricula (Federal Ministry of Education, Science and Research, 2017) and teacher practices and in addition to introducing the new concept of MU, as already touched upon in section 4.1. The stages of each lesson are based on the 12 basis models by Oser and Baeriswyl (2001) and general outlines of deep and surface structure are summarised in Tab. 3 and 4.

Tab. 3. General lesson outline for first session according to basis model

Session I - manufacturing a ruler **Basis model: 'learning through personal experience' (Oser & Baeriswyl, 2001)**		
time	**deep structure**	**surface structure (outline of visible activities)**
10 min	planning of action	introductory text on MU (faster than light measurement at CERN?) with follow up discussion on NOS and MU, planning of action: learners manufacture a measuring tool to find out more about them
15 min	performance of action	learners build their own inch ruler (scaffolding worksheet)
15 min	construction of meaning	rating of trustworthiness and comparison with further tools that measure length
	generalisation of the experience	introduction of TBU, which applies for all types of gauges
10 min	reflecting on similar experiences	online survey on examples with general relevance to TBU

The first lesson's structure is derived from the model "learning trough personal experience" (Oser & Baeriswyl, 2001) and its core elements are learners building their own measuring tool and the introduction of TBU. An introductory text on the measurement of superluminal speed at CERN in 2011 and the resulting violation of special relativity in the context of MU was developed to introduce the term MU. Afterwards learners are split into two groups and have to collect ideas and arguments for two opposing opinions on NOS. It is followed by a plenary discussion led by the teacher and students have to argue for either opinion A (i.e., *physics is an accurate field of science*) or opinion B (i.e., *No scientific knowledge can be 100% accurate*). An essential aspect at this stage is that leaners are not provided with a right or wrong answer at this point. Rather is this instance framed as a starting point for the upcoming lessons to find out more about MU to be able to answer this question with more certainty. Hence, this starting discussion and students' beliefs and ideas can be reviewed at a later stage of the TLS.

Tab. 4. General lesson outline for second session according to basis model

Session II – measuring length and interpreting results **Basis model: 'concept building' (Oser & Baeriswyl, 2001)**		
time	**deep structure**	**surface structure (outline of visible activities)**
5 min	activation of pre-knowledge	summary and reflection of previous session in plenary (schematic knowledge), teacher states that the objective for this session is to 'test' the self-made inch rulers for measuring length
25 min	introduction and execution of a prototype (new concept)	pair work: measuring length of papers strips and subsequent calculation of an unknow total length, data from the whole class is collected (modified strips ensure normal distribution of data) statistical analysis in class with introduction of TBA (e.g. teacher-centred, digital tools, individually, etc.)
10 min	development of characteristics	mean and MU are related to trustworthiness of a measurement result (advanced tasks on worksheet or online tool e.g. socrative, nearpod, schoology, etc.)
10 min	application and transfer of the concept	individual analysis of a given set of data including correct reporting of the result (mean ± uncertainty)

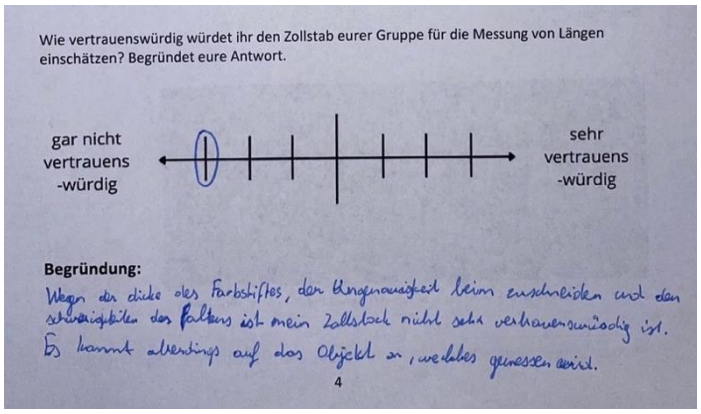

Fig. 2. Rating scale for trustworthiness (cf. Loidl, 2021). Reasons for ratings are thickness of the felt pen, aspects of cutting the paper and difficulty of folding. The learner adds that rating depends on the object that should be measured with the ruler.

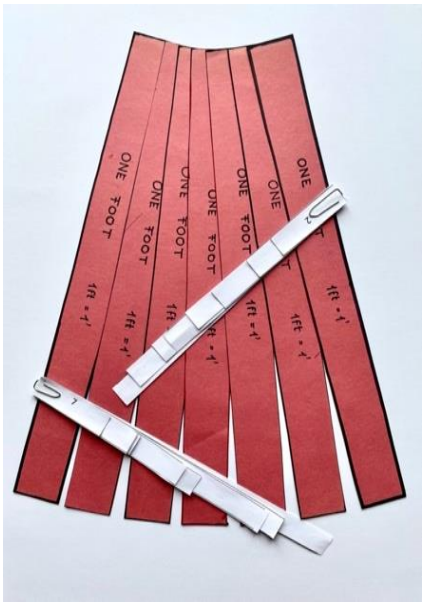

Fig. 3. Cardboard calibration standards and numbered paper strips

To find out more about MU and measurement tools in general, a step-by-step guided task to manufacture an inch ruler in pairs follows. Especially in countries outside of the Anglosphere, where the imperial system has no to little curricular tradition or everyday relevance, this adds aspects of novelty in opposition to merely building a ruler that is already familiar to learners. The steps of this task correspond with the first two subtypes of TBU, when leaners have to transfer the length of a calibration standard of one foot onto a piece of paper (uncertainty of calibration) and fold a second strip of paper with the same length into 12 parts (uncertainty of linearity). The calibration standard can either be provided with a piece of wood or alternatively strips of cardboard (Fig 3). Apart from that, the materials needed can be seen in Fig. 5 and align with the design principle of practicability. The last step of this task includes the estimation of the trustworthiness of each individual ruler using a rating scale with explanations (Fig. 2) that has been adopted from Loidl (2021). Following activities included the definition of TBU, different notations of MU, comparison of different measurement tools (triangle ruler $\measuredangle l = \pm 1$ mm and tape measure $\measuredangle l = \pm 1/16$ "), a fact box on TBU and further differentiated tasks

The second lesson resumes with the handling of the inch ruler from lesson one with an TBU of $\pm 1''$ for measuring length and follows the stages of the basis model "concept building" (Oser & Baeriswyl, 2001). Learners work in pairs and are given paper strips of a specific length (40″) that has been previously measured and cut with tools of a high "trustworthiness" (i.e., small MU). Learners complete a table with the values for the individual strips to calculate the sum and therefore total length of the paper before it was cut. Each group gets numbered strips because the cutting of the 40 ″ lengths has been done randomly before so that each group measures different strip lengths. The learners are given the Information that the total length for all groups is the same. However, its value must not be known by the leaners. During this phase learners will engage with aspects of reading uncertainty as the task and the teacher will ask them to provide values only as accurate as it is possible due to the scale of their inch ruler. If students indicate values with a decimal, they are asked to round to the nearest whole number. This activity again involves the estimation of trustworthiness not of the utilised tool but the determined value itself with the same rating scale that has been used in the previous lesson.

Different values for the total length are gathered and documented in class afterwards (Fig. 4). The paper strips are modified in a way that outliers or a at least a spread of data that resembles Gaussian distribution can be realised. Gaussian distribution was chosen due to it being referred to as the most common distribution according to GUM (BIPM, 2008, p. 24) and relevant passages in the Austrian national curriculum for mathematics (Federal Ministry of Education, Science and Research, 2017). Additionally, the concepts of the true value, standard deviation of the mean and a Gaussian distribution are conceptionally linked within that Gaussian model and were considered complementary to the concept of MU on the basis of MER (cf. "elementaria" Kattmann et al., 1997). It should be noted that Gaussian distribution and actual calculation of standard deviation of the main do not represent such existential content as the true value and MU. That is why the two are also not included in defined core concepts which are introduced in the previous section. Regarding the complexity of a Gaussian distribution, Fig. 4 can simply be described as a graph that provides information about an area where the true value can be found with a certain likelihood in class (e.g. the peak of the graph shows an area were the true value is very likely). This approach has been included explicitly in the didactic commentary. Engaging the Gaussian model on a deeper level can be considered optional and as context dependent (e.g. school type, implementation of mathematical tools, variable mathematical competences, etc.). Follow-up questions ask for the measurement value that is most likely closest to the true value and reasons for the spread of data. It is only after these activities that it is revealed to the learners that all groups were given a total length of 40″ measured with means of high trustworthiness, which is also used as a transition to the next phase of the lesson. This part is the statistical analysis of the data collected in class to calculate the mean and standard deviation of the mean to be able to provide a measurement value with the appropriate MU. Different methods of conducting this analysis (e.g. teacher-centred, application of digital tools, etc.) can be chosen individually by the teacher. Concluding tasks involve correct reporting of results, ranking and comparing results according to their trustworthiness and the comparability of different results based on MU.

The sum of all materials including lesson plans, worksheets and a didactic commentary for teachers can be accessed via the following link: https://aeccp.univie.ac.at/lehrer-innen/fuer-den-unterricht/

Fig. 4. Example for in class documentation of measurement values for total length of a paper strip

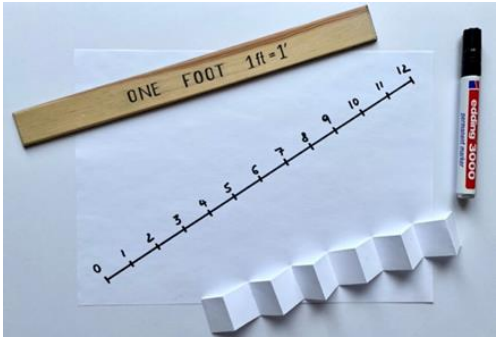

Fig. 5. Materials for manufacturing an inch ruler with $\Delta l = \pm 1$"

4.3 Significant results of all cycles

4.3.1 Expert interviews

The cycles before probing acceptances (DC II) were conducted, were dedicated to the development of design principles, definition and selection of core concepts (e.g. MU, true value, mean, etc.) and choice of models for the TLS. This stage led to the results presented in the previous section (e.g. key ideas and learning goal, design principles, structure and tasks of the TLS). This process involved many steps that cannot be displayed in their entirety by this article. Ideas on the concept of uncertainty propagation, for instance, were not further developed due to issues of learners' mathematic skills at the targeted age group, just to name one aspect that had been discussed during this expert interview stage. Testing of the experimental setting (Micro-DC) showed no difficulties when learners carried out the experiments. Testing took place in two different classes and the teacher was provided with an observation form for post-teaching reflection as a basis for further discussion. During this discussion, no issues regarding execution, difficulty or applicability were raised. On the basis of expert interviews, two hypotheses were formulated for the cycle of probing acceptances. Additional questions and rating scales were added to probing guideline questions relating to motivational aspects of learners and the experimental step of folding a strip of paper into twelve pieces.

4.3.2 Probing acceptances

These were conducted with one pair of students as well as one single student from the same class and audio recording were taken for post probing analysis. 8 categories covering reasoning processes, solution strategies and utterances and activities leading to incorrect assumption or solutions were defined for the analysis of the recordings. Learners did not show any difficulties regarding task comprehension and during the experimental settings. Key concepts were also defined and distinguished correctly in post-probing debriefings. Ratings of trustworthiness were in general low for both the manufactured ruler and learners' measurement results with it. However, learners named the provided materials and methods used (e.g. folding, using a thick felt pen, etc.) as reasons for a low rating instead factors related to the experimenters (faulty human execution of a task). One learner articulated this for example by saying that she sees the aim of building the ruler with such *simple tools* and additionally transfers this issue and therefore the concept of MU to a more general context according to the following statements (Fig.7):

> *Well, this [MU] can be applied to other measuring tools as well.*
> *Even triangle rulers are not 100% precise because their marks also have some thickness to them and probably also the machines that produce such rulers don´t work absolutely precise.*

Fig. 7. Learner transcript from probing acceptances

Furthermore, gaining knowledge about imperial units and learning how to fold something into 12 parts was mentioned explicitly by all participants as personally valuable to them. The concept of random deviation as one reason for MU and the concept of randomness in general seemed to be at least not entirely relatable for these participants. The wording trustworthiness had been discussed in detail with all participants. Despite initial connotations of trustworthiness with people instead of measurements (e.g. a trustworthy person) learners used this terminology consistently throughout the probing phase and additionally considered it the most suitable in this specific context after a short reflection and discussion. Surprisingly, the term error had not been mentioned by any student during this stage. As far as the hypotheses based on the previous cycles are concerned, Likert scale rating (1-5) showed that the topic of MU in general was considered useful and interesting by the participants and there were no indications of negative motivational impact (average rating of 3.5). Learners also followed the step-by-step instruction when building their inch ruler and did not think of further scaling by additional folding. The wording of the task on paper strip measurements also ensured that values were not provided using decimals. Additionally, all tasks on ordering measurement results according to trustworthiness and correct reporting of results were solved correctly in with all probing participants.

4.3.3 Testing in classroom settings

These cycles involved collection of data via multiple tools. The analysis of data provided by video material was performed using the same categories used during probing acceptances. Categories were both predefined but also adapted and expanded during this stage. The teaching materials were collected after each lesson for analysis of worksheet exercises and tasks, in addition to data on additional tasks collected via an online tool in class. The final stage of these cycles consisted of data collection using the developed written assessment. Post-teaching discussion based on two different observation forms with the class teachers, who were observing throughout the whole lessons, were conducted, corresponding with the qualitative criterion of "communicative validation" (Mayring, 2016, p. 45). Evaluation took place in a total number of three classes (age 16-17) and two different schools in Austria. The total number of learners that participated in these testing settings was 32, with some being absent at some sessions. An extensive scale

analysis and categorisation resembled the findings from the probing acceptances. No incidents of misunderstanding or lack of comprehension regarding wording and tasks could be observed. Usage of the term MU could already be observed at the start of all lessons during the classroom discussion on NOS. Various views on NOS were detected from the idea that nothing is verifiable and knowledge about quantum theory to physics in its nature being 100 % precise with a lack of human expertise until the current time period (e.g. physics is accurate, but our instruments and models are not). Examples of correspond statements by leaners are provided on the next page in Fig. 8, 9 and 10.

> *Nobody said that physics is NOT precise/accurate. We just talked about that physics is not 100 % precise/accurate.*
> *Yes, but science is manmade, so it is based on humans doing science and we know that humans are not perfect and therefore science cannot be perfect either.*
> *Those measurement uncertainties of course make an Experiment inaccurate, which could be seen in this CERN experiment.*
> *Oh yes of course we have to distinguish between theory and practice […] if we come up with a scientific law and it can be proven, then it is possible to be accurate in theory.*

Fig. 8. Learner transcripts from testing in classroom setting during discussion on NOS and MU

Utterances during a group discussion are provided in Fig. 9:

> *L1: And science itself consists of formulas and those are accurate, or they exist […] in a sense of physics exists and it is accurate, and it cannot be changed. It can only be the case that you don´t have any results yet. Like with the sun, which also goes up every day.*
> *L2: Yes, I would have said that the mistake does not lie with science but with [pause and both students simultaneously] THE MEASUREMENTS!*
> *L1: But are measurements part of physics [in contrast to being part of human application]?*
> *L1 in plenary: We agreed that there are mistakes are not part of science but more part of measurements, so the people that take those measurements and that is why inaccuracy (germ. Ungenauigkeiten) occur. But theoretically Science itself is precise.*

Fig. 9. Learner transcripts from testing in classroom setting during pair discussion on NOS and MU

The analysis also showed that there were only two instances of learners using the term error (germ. *Messfehler*), when referring to errors made by humans such as messy executions and no other references to errors were recorded elsewhere.
Ratings of trustworthiness, were again, predominantly on the negative scale for the participants. Nonetheless, even those who chose higher ratings provided the same arguments (e.g. folding of paper, scale of the ruler, thickness of scissor blades and pens, etc.). Examples for comments by learners regarding this issue during the task are:

> *But the thick pen is not precise, is it?*
> *The pen is thicker than the crease, that´s a problem!*
> *But we should cut out strips inside of the line, right?*
> *Folding into thirds cannot be totally accurate!*

Fig. 10. Learner transcripts from testing in classroom setting on reasons for ratings of trustworthiness

Relatively high ratings of trustworthiness also referred to these methods and tools and related them to the purpose of length measurement (e.g. it can be considered trustworthy for the goal of measuring paper strips but not for anything that requires more precision). Errors in the sense of GUM during the various processes were not referred to. A question item during the lessons on reasons for deviation of measurement values showed difficulty regarding the concept of random deviation. Hardly any learners saw this as a reason, which had already been the case during probing acceptances. MER and previous cycles resulted in the conclusion that this concept however is not essential for the understanding of MU. The wording coincidence or randomness (germ. *der Zufall)* seemed to be more inappropriate or unrelatable for learners as random deviation (germ. *zufällige Abweichung*) was selected more often in another item. This could also indicate why such reasoning was not used by learners when rating trustworthiness. Nonetheless, an answer stating that every series of measurements is characterised by deviation of some degree, was selected by over 75% of all participants. Statistical analysis did not show to be too demanding for the sum of all leaners. Especially calculation of the mean was achieved throughout all classes individually, whereas determination of the deviation of the mean was solemnly achieved via calculators in one class and online tools or Excel in the two other classes. This stage lasted for a maximum of 10 minutes with all participants and class teachers also explicitly recorded that the mathematical level was appropriate.

In class online tasks on ranking and comparing results according to their trustworthiness showed striking results. In one class a ranking task was successfully accomplished by 12/12 learners and the other two classes showed equal achievements. Another task on the correct reporting of results also shows that hardly any answer showed selection of a result without an uncertainty (cf. answer C, Fig. 12). Analysis of collected worksheet highlights this circumstance as no answers without documentation of MU were found. However, it cannot be said whether these answers were completed after in class discussion of solutions. Summarised results for online tasks that correspond with specific learning objectives are illustrated in Fig. 11 and 12. Fig. 11 shows that 85 % of all participants were able to order measurements values correctly in class, with answer A using providing a solution with the same order starting with highest trustworthiness instead of lowest. In Fig. 12 answer B and E were false options due to incorrect rounding for B and a missing indication of a unit in E.

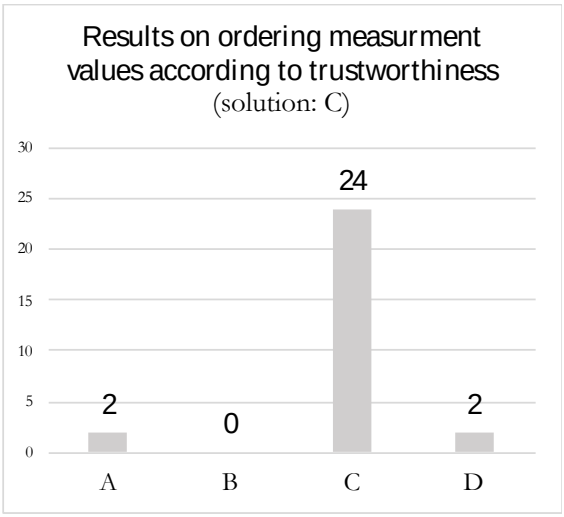

Fig. 11. Summarised results for in class item on evaluating trustworthiness of results

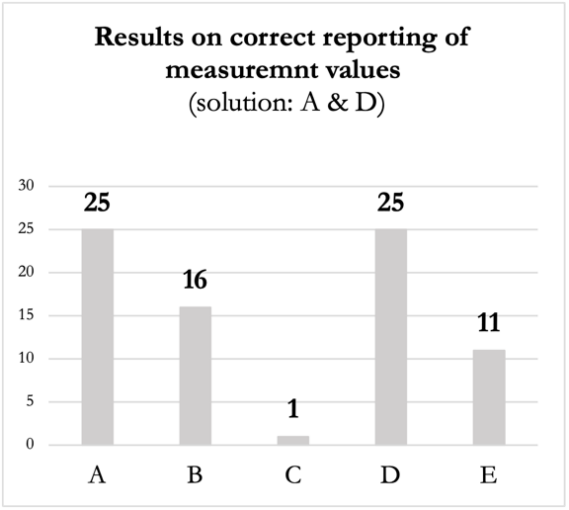

Fig. 12. Summarised results for in class item on correct reporting of results

In total, 64 % of all answers provided in this multiple-choice item were correct and a striking result is that reporting without MU was only chosen once in all three evaluation settings. Results for two items from the written assessment show equally successful student performances. An item on ordering according to trustworthiness was solved correctly by 24/24 learners. Analysis of collected worksheets additionally showed that comparison of trustworthiness between a triangle ruler and a 1/16 ", as a transfer exercise (application of new knowledge in a different context), was accomplished by almost all learners. Results on correct reporting are even more significant with this testing instrument, as correct answers were given 23/24 times.

Summarised results for each individual learning goal and achievement according to each relevant testing instrument are shown in Tab. 3. Indication of hyphen expresses that validation for this specific learning goal was not possible due to the nature and design of that chosen instrument. However, this does not indicate that learning goal has not been achieved. A benchmark of 75 % correct answers was set for given achievement of objectives by the instruments of in class online tasks and the written assessment. A slightly refined version of the written assessment was used for the last testing cycle, which is why results for two items were removed from the evaluation results (cf. BRG 14, Tab. 3). Written assessment items were designed in way that leaners could either acquire full or no points.

As outlined in section 2.1.2 research on students' beliefs on MU is still quite limited. One issue that was observed during probing acceptances and all classroom testings was that learners would initially always provide the smaller MU, after determining TAU and TBU, as it was considered 'better'. However, onetime clarification that providing a smaller MU than the TBU of a gauge for instance would portray 'fake certainty' (germ. *falsche Vertrauenwürdigkeit*), subsequent tasks were all solved with the appropriate MU.

Tab. 5. Summary of learning objectives and fulfilment according to each testing instrument

learning goal	probing acceptances	video data analysis & observation forms	written assessment		in class online tasks
			BRG 14	BRG Lienz	
… **describe** steps during the production and use of a measuring tool that lead to TBU.	✓	✓	-	✓	✓
… **evaluate** measurements based on their trustworthiness. They can then **rank** different measurements based on their MU and **make comparisons** regarding trustworthiness.	✓	✓	✓[4]	✓	✓
… **correctly report** measurement results (mean ± uncertainty) and distinguish between correct and incorrect indication of scientific measurement values.	✓	✓	✓	✓	✓
… **calculate** the mean and standard deviation of the mean for a given set of measurements.	-	✓	✓	✓	-
… **differentiate** between TAU and TBU.	✓[5]	✓	-	✓	-

4.3.4 Expert discussion

Methodology and design of the TLS had been approved during this expert discussion that involved researchers and concept developers from the Austrian Educational Competence Centre Physics (AECC-Physics) at the University of Vienna. It was agreed that it is vital to the TLS that learners do not end up with the conception that physics can never be 100 % accurate and therefore science cannot be seen as trustworthy. This resulted in an explicit commentary in a teacher's guide that has been developed for the teacher materials. It is mentioned that teachers should keep this issue in mind and refer to the initial discussion of the first lesson repeatedly when the concept of trustworthiness and determining factors for highly trustworthy scientific results are established in class.
Discussions between experts on the terminology of trustworthiness resembled reflections and thought processes from early stages of the DBR project. Alternative wordings depicting the concept of MU that have been mentioned and discussed were *credibility, accuracy, precision, reliability*, germ. *Zuverlässigkeit*, germ. *Validität* germ. *Aussagekraft*, and germ. *Genauigkeit*. It should be noted that these terms were discussed within the context of alternatives to the term trustworthiness and stem from a rather vernacular or spoken language terminology instead of technical terms. Experts agreed that additional research on learners' associations with such terms is needed to investigate which wording is most suitable to manifest the right conceptual understanding for learners. However, terms such as reliability or validity have been evaluated as problematic due to their specific definitions in the field of science (Altrichter et al., 2018, p. 104). It was argued that, at least within a German speaking context, trustworthiness (germ. *Vertrauenwürdigkeit*) accounted for high learner acceptance during probing and also lead to achievement of learning objectives that included trustworthiness according to the data collected via quantitative testing instruments. Moreover, trustworthiness was described suitable, as all three aspects of qualitative criteria are covered referring to aspects of the 'trustworthiness of a normal distribution model', which would not be given by a word such as germ. *Zuverlässigkeit* (e.g. 'that would not be *zuverlässig*, but within the boundaries of the model *vertrauenswürdig*). Conceptional value of statistical analysis according to GUM was also part of a stimulating discussion. Consensus was reached that learners should be provided with opportunities to apply new concepts and that calculation and especially interpretation of the deviation of the mean can be such an opportunity. The key conceptional difference between standard deviation and deviation of the mean due to its relation to the true value was also highlighted during this stage of the discussion. Nonetheless, it was agreed on that mere calculations do not add conceptional value to curricula in general. Still, that corresponds with the intentional goal of such items in this TLS in the sense that focus was put on stages after the statistical analysis based on its results.

[4] This was the only item where the benchmark score was just slightly exceeded. Scores for all other items where continuously high and for some items even higher than 95 %.
[5] Differentiation between the two concepts was regarded as given if learners were able to successfully define TBA and TBU and explain the two concepts in their own words.

6 Conclusion

Considering the underlying research question of this project, the data obtained during the DBR - cycles (Tab. 5) expresses that all defined learning objectives were achieved for this TLS on MU. Nonetheless this has to be understood within the limitations of a DBR approach elaborated in section 3.1. Especially the aspect of generalization should be seen critically as this is deemed problematic within DBR in general. Regarding a rather limited number of participants during cycles that involved the testing of the TLS in class, additional cycles that contribute insights on a broader scale have to be considered. Results of probing acceptances and findings during expert discussions corresponded with the idea that the TLS has the potential to be successfully applied within different contexts as well. Based on the nature of DBR, the clinical data gathered during this project indicates that the TLS has at least similar potential in different settings despite aspects of humble or local theories in DBR (Guisasola et al. (2023, p. 23). The "local results" showed that learners were able to successfully assess the trustworthiness of scientific measurement practices and results, as well as compare such based on what was framed 'trustworthiness' in this teaching concept. It can be argued that this is an indicator for conceptual understanding of MU and its application in various context. Such application also involved knowledge about correct use of scientific formalism required for reporting MU and determination of TBA and TBU.
Observations showed that the concept of MU was used and referred to by learners from an early stage of this setting. The same accounts for wording from the semantic field of trustworthiness (e.g. *Vertrauenswürdigkeit, sehr vertrauenswürdig, unvertrauenswürdig, vertrauenswürdiger*, etc.) and instances referring to errors were almost non-existent. It might be argued that no items on the conceptual distinction between error and MU were implemented in the TLS. However, the concept of fundamental importance and relevance for scientific practice is MU and the TLS successfully draws attention to it. Types of errors are yet defined via a footnote in the worksheet packet to provide clear definitions and differences between the two. This implies that it is not possible to completely dismiss the idea of learners blending concepts related to error and MU within that context. Affirmation of such a distinction between a concept that would correspond with the technical term of error and MU could be detected in further experimental settings. If learners would record and present measurement values without any additional information, this would indicate that the concept of MU has not been fully developed or distinguished from error. In general, these materials represent a first introduction to MU that calls for repeated practice and revision of competences and knowledge regarding MU. Every subsequent experimental setting should be seen as an opportunity to draw learners' attention towards MU again. It can be argued that the TLS introduces a relatively high number of new technical terms, such as true value, MU (Type A and B), and measurement value and mean. However, what can be considered of fundamental importance is the key idea that a measurement value should always be indicated with the appropriate MU. Learners considering MU during other experiments would be a main additional indicator for achievements of such objectives. In essence, the key point might not be whether learners grasp how to report a value with its corresponding MU. Instead, it could simply revolve around the necessity of providing MU because without it, a measured value is meaningless and becomes unnecessary. Technical issues, such as rounding and suitable statistical analysis can be considered skills rather than understanding of MU, which can and also should be revised at any possible occasion within multifaceted scientific settings.
Looking at learner's beliefs about physics, the TLS enables learners to associate MU with NOS and as a consequence also with scientific methods and tools instead of what could be termed 'human errors'. Evidence for this is suggested by provided reasoning for high or low trustworthiness during rating tasks. Learners predominantly referred to factors of the experimental setting instead of bad execution or 'human factors' (e.g. I am not good at experimenting, I did not fold the paper properly, etc.) when providing justifications for their ratings. That is why the activity of building your own measuring tool can be considered to represent a fitting activity to introduce the concept of MU (cf. learning trough personal experience, Oser & Baeriswyl, 2001). It might be the case that erroneous human actions were not reported as relevant due to rather low complexity of the task, when building an inch ruler. However, observations in class showed that folding into thirds to achieve 12 equal parts posed a challenge for the majority of learners and low complexity would contradict mostly low ratings of trustworthiness for the ruler.
Difficulties in comprehension could not be detected with standard deviation of the mean and the true value but with the concept of randomness. Despite 'random deviation' not appearing to be that problematic for learners, the wording of randomness or chance (germ. *Zufall*) emerged as an unrelatable idea. Whether this phenomenon is more likely related to semantic issues or situated on a deeper level could not be evaluated in this setting. Similarly, further research is needed on the term trustworthiness and learners' conceptual associations with it. Both aspects mentioned are currently investigated in research projects at the University of Vienna.
In general, it is essential to take into account the perspective presented, within the context of DBR by McKenney & Reeves (2018, p. 21). It is suggested that when considering a contextual setting, one should regard generalizations as working hypotheses rather than definitive conclusions. This role of the context puts the sum of results and conclusions into a different perspective. Replication of interventions entails obstacles "since there are so many factors at play when interventions go live" (McKenney & Reeves, 2018, p. 20). This is relevant when reflecting on results or implication of findings during such a project, but also when the adaption of TLS materials by teachers is discussed. Adaption of materials and varying performances by practitioners are another substantial element that makes assumptions on generalisations, such as beneficial integration of the concept of true value or the positive effect of trustworthiness, on a broad scale at least challenging. Additionally, a strong indicator for lasting comprehension of the concept of MU would be if learners consider MU during experimental settings at later stages. This factor was not investigated within this setting. Furthermore, definition of key ideas and learning objectives is backed up by hypotheses based on research

within the scientific community and MER and not by investigation that involved collection of data in the field (cf. "ecological validation" Haagen-Schützenhöfer & Hopf, 2020, p. 4).
Ultimately, the main relevance of MU for science education can be seen in its impact on learners' ideas on NOS. Even though positive effects on deeper understanding of NOS were not under investigation in this setting, video analysis hints towards application of MU when referring to scientific models and methods in general during classroom discussions. MU and NOS have, as argued for in section 1.1., educational potential that has not been fully explored yet. This is why extended study on this relation can also only be of interest in science education research. The implementation of MU using the materials provided for this TLS can also facilitate this type of ongoing research. The required materials can be found on the website of the AECC-Vienna[6]. However, it should be noted that for the time being these are only available in German as issues of terminology would require more than a mere translation but also MER of applicable wordings for an English-speaking context.

References

Altrichter, H., Posch, P., & Spann, H. (2018). *Lehrerinnen und Lehrer erforschen ihren Unterricht* (5th ed.). Verlag Julius Klinkhardt UTB.

Anjobi Videoarchiv. (2023, July 27). 2023-07-27 1045 Salzburger Festspiele ANJOBI_Cut [Video file]. YouTube. https://www.youtube.com/watch?v=6m4iFg5I4bk

Barab, & Squire, K. (2004). Design-based research: putting a stake in the ground. *The Journal of the Learning Sciences, 13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301_1

Bärenthaler-Pachner, R. (2022). Entwicklung und Evaluation einer Lernumgebung zum Thema Messunsicherheit in der Sekundarstufe II. [Masterarbeit, Universität Wien]. Universität Wien.

BIPM (2008). *JCGM 100:2008 Evaluation of measurement Data - Guide to the Expression of Uncertainty in Measurement* (GUM). [International Organization for Standardization, Geneva, 1995, ISBN 92-67-10188-9].

Boczianowski, F. & Kok, K. (2020). Modelle empirisch prüfen - Frequenzmessung an stehenden akustischen Wellen mit dem Smartphone. *MNU Journal, 73*(4), 295-299.

Bogner, A., Littig, B. & Menz, W. (2009). Introduction: Expert interviews - An introduction to a new methodological Debate. In: Bogner, A., Littig, B. & Menz, W. (Eds.), *Interviewing Experts. Research Methods Series.* Palgrave Macmillan, London. https://doi.org/10.1057/9780230244276_1

Brown, A. (1992). Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141–178. doi: 10.1207/s15327809jls0202_2

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

Buffler, A., Allie, S., Lubben, F., & Campbell, B. (2005). *Introduction to measurement in the physics laboratory. A probabilistic approach.* https://www.nhn.ou.edu/~strauss/Quarknet/PDF_Curriculum/Intro_To_Measurment.pdf

Buffler, A., Lubben, F. & Bashirah, I. (2009). The relationship between students' views of the nature of science and their views of the nature of scientific measurement. *International Journal of Science education, 31*(9), 1137-1156. https://doi.org/10.1080/09500690802189807

Buffler, S. & Lubben, F. (2008). Teaching measurement and uncertainty the GUM way. *The Physics Teacher, 46*(9), 539-543.

Burton, D. (2015, April 19). *The essence of science in 60 seconds (Richard Feynman)* [video] YouTube. https://www.youtube.com/watch?v=LIxvQMhttq4

Cobb, P., Confrey, J., Di Sessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher, 32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Confrey, J. (2005). The evolution of design studies as methodology. In *The Cambridge Handbook of the Learning Sciences* (pp. 135–152). United Kingdom: Cambridge University Press. doi: 10.1017/CBO9780511816833.010

Covitt, B. A., & Anderson, C. W. (2022). Untangling trustworthiness and uncertainty in science: Implications for science education. *Science & Education, 31*(5), 1155 - 1180. https://doi.org/10.1007/s11191-022-00322-6

Department for Education - Canton Bern. (2017). *Curriculum 17. Lehrplan 17 für den gymnasialen Bildungsgang.* https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjdiqrjibuBAxWi2gIHHdPjD4sQFnoECBwQAQ&url=https%3A%2F%2Fwww.bkd.be.ch%2Fcontent%2Fdam%2Fbkd%2Fdokumente%2Fde%2Fthemen%2Fbildung%2Fmittelschulen%2Fgymnasium%2Famsgym-lehrplan-17-gesamtdokument.pdf&usg=AOvVaw2sAXf9cJDT7r4hQoKVRiED

Department for Education - Canton St. Gallen. (2019). *Physics curriculum for upper secondary in the canton of St. Gallen.* https://www.sg.ch/bildung-sport/mittelschule/lehrplaene-und-stundentafeln/gymnasium/_jcr_con-

[6] https://aeccp.univie.ac.at/lehrer-innen/

tent/Par/sgch_accordion_list/AccordionListPar/sgch_accordion/AccordionPar/sgch_downloadlist/DownloadListPar/sgch_download.ocFile/Lehrplan%20MAR%20200810%20Inhaltsverzeichnis%20angepasst%20Oktober14.pdf

Department for Education of the UK. (2014). *GCE AS and A level subject content for biology, chemistry, physics and psychology.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/593849/Science_AS_and_level_formatted.pdf

Department for Education of the UK. (2015). *National curriculum in England: science programmes of study (until key stage 4)* https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

der Humboldt-Universität zu Berlin der Humboldt-Universität zu Berlin]. https://edoc.hu-berlin.de/bitstream/handle/18452/25656/dissertation_kok_karel.pdf?sequence=6

Duit, R. & Treagust, D.R. (2003). Conceptual change: a powerful framework for improving science teaching and learning, *International Journal of Science Education, 25*(6), 671-688. https://doi.org/10.1080/09500690305016

Duit, R., Groppengießer, H., Kattmann, U., Komorek, M. & Parchmann, I. (2012). The model of educational reconstruction – a framework for improving teaching and learning science. In D. Jorde & J. Dillon (Eds.), *Science education research and practice in Europe* (13-37). Springer.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *The Journal of the Learning Sciences, 11*(1), 105–121. doi: 10.1207/S15327809JLS1101_4

Evangelinos, D., Psillos, D., & Valassiades, O. (2002). An investigation of teaching and learning about measurement data and their treatment in the introductory physics laboratory. In D. Psillos & H. Niederrer (Eds.), *Teaching and learning in the science laboratory* (179- 190). Kluwer Academic.

Federal Instutuation for Education Standards and Research in Education – Munich. (2023). *Physciscs Curriculum for Gymnasium/Upper Secondary.* https://www.lehrplanplus.bayern.de/fachlehrplan/gymnasium/13/physik/grundlegend

Federal Ministry of Education, Science and Research. (2017). *Austrian national curriculum for general secondary education.* https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568&FassungVom=2017-08-31

Feulner, B., Ohl, U. & Hörnmann, I. (2015). Design-Based Research - ein Ansatz empirischer Forschung und seine Potentiale für die Geographiedidaktik. *Zeitschrift für Geographiedidaktik, 43*(3), 205-231.

Fussell, R. & Holmes, N. (2022). *Machine learning for automated content analysis: characteristics of training data impact reliability.* Paper presented at Physics Education Research Conference 2022, Grand Rapids, MI. Retrieved October 8, 2023, from https://www.compadre.org/Repository/document/ServeFile.cfm?ID=16231&DocID=5600

Garrett, J., Horn, A. & Tomlinson, J. (2000). Misconceptions about error. *University Chemistry Education, 4*(2), 54-57.

Glomski, J. & Priemer, B. (2010). *Modellierung eines adäquaten Umgangs mit Messunsicherheiten.* PhyDid B - Didaktik Der Physik - Beiträge Zur DPG-Frühjahrstagung. https://ojs.dpg-physik.de/index.php/phydid-b/article/view/141

Gränicher, W. (1996) *Messung beendet – was nun? Einführung und Nachschlagewerk für die Planung und Auswertung von Messungen* (2nd ed.). Vdf Hochschulverlag AG-ETH Zürich.

Gräsel, C. & Parchmann, I. (2004). Implementationsforschung - oder: Der steinige Weg, Unterricht zu verändern. *Unterrichtswissenschaft, 32*(3), 196-214.

Guisasola, J., Zuza, K., Sarriugarte, P. & Ametller, J. (2023). Research-based teaching-learning sequences in physics education: A rising line of research. In M. F. Taşar and P. R. L. Heron (Eds.) *The International Handbook of Physics Education Research: Special Topics.* (26-1–26-26). New York: AIP Publishing.

Haagen-Schützenhöfer, & Hopf, M. (2020). Design-based research as a model for systematic curriculum development: the example of a curriculum for introductory optics. physical review. *Physics Education Research, 16*(2), 020152. https://doi.org/10.1103/PhysRevPhysEducRes.16.020152

Haagen-Schützenhöfer, C. (2015). *Kumulative Habilitationsschrift mit dem Schwerpunkt Lehr- und Lernprozesse im Anfangsoptikunterricht der Sekundarstufe I* [Habilitationsschrift, Universität Wien]. Universität Wien.

Harlen, W. (1999). *Effective teaching of science: a review of research*. Scottish Council for Research in Education.

Heinicke, S. (2012). *Aus Fehlern wird man klug. Eine genetisch-didaktische Rekonstruktion des „Messfehlers".* [Dissertation, Universität Oldenburg]. Berlin: Logos Verlag.

Heinicke, S. (2014). Experimentieren geht nicht ohne (Mess-)Unsicherheiten. *Naturwissenschaft im Unterricht Physik: Experimentieren Gestalten,* (144), 29–31.

Heinicke, S. & Holz, C. (2018). Mit Messfehlern umgehen und Messungen evaluieren. Neue Wege der Fehlerbetrachtung am Beispiel der e/m-Bestimmung. *Naturwissenschaft im Unterricht Physik, 29*(168), 18-23.

Heinicke, S. & Holz, C. (2019). Messunsicherheit - ein ungeliebter Gast im Physikunterricht? In C. Maurer (Ed.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe* (39th ed. 89-92). Gesellschaft für Didaktik der Chemie und Physik. https://doi.org/10.25656/01:16753

Heinicke, S. & Holz, C. (2020). Messfehler 2.0. *Naturwissenschaften im Unterricht Physik, 177/178,* 33–38.

Heinicke, S. & Riess, F. (2009). How to cope with Gauss's errors. In G. Cakmakci & F. Tasar (Eds.), *Contemporary science education research: learning and assessment.* (23-32). ESERA.

Heinicke, S., Glomski, J, Priemer, B. & Rieß, F. (2010). Aus Fehlern wird man klug – Über die Relevanz eines adäquaten Verständnisses von „Messfehlern" im Physikunterricht. *Praxis der Naturwissenschaften - Physik in der Schule, 59*(5), 26-33.

Hellwig, J. (2012). *Messunsicherheiten verstehen - Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik.* [Dissertation, Ruhr-Universität Bochum). Ruhr-Universität Bochum, Universitätsbibliothek. https://d-nb.info/1221367757/34

Hellwig, J. & Heinicke, S. (2020): Messfehler – wann, warum und wie? *Naturwissenschaften im Unterricht - Physik 177/178*, 28-32.

Hellwig, J., Schulz, J. & Priemer, B. (2017). Messunsicherheiten im Unterricht thematisieren - ausgewählte Beispiele für die Praxis. *Praxis der Naturwissenschaften – Physik in der Schule 66*(2), 16-22.

Höttecke, D. (2013). *A sketch of the problem of authentic inquiry-based learning from a history of science perspective.* [Paper presented at the Twelfth International History, Philosophy, Sociology & Science Teaching Conference (IHPST)]

Hu, D. & Zwickl, B. M. (2018). Examining students' views about validity of experiments: From introductory to Ph.D. students. *Physical Review. Physics Education Research, 14*(1), 010121. doi: 10.1103/PhysRevPhysEducRes.14.010121

Jung, W. (1992). Probing acceptance: a technique for investigating learning difficulties. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: theoretical issues and empirical studies* (278-295). IPN.

Kattmann, U., Duit, R., Groppengießer, H. & Komorek, M. (1997). Das Modell der Didaktischen Rekonstruktion - Ein Rahmen für naturwissenschaftsdidaktische Forschung und Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften, 3*(3), 3-18.

Klüver, J. (1979). Kommunikative Validierung - einige vorbereitete Bemerkungen zum Projekt Lebensweltanalyse von Fernstudenten. In T. Heinze (Ed.), *Theoretische und methodologische Überlegungen zum Typus hermeneutisch-lebensgeschichtlicher Forschung* (69–84). Werkstattbericht Universität Hagen.

Kok, K. (2022). *Certain about uncertainty. What students need to know about measurement uncertainties to compare data sets.* [Dissertation, Mathematisch-Naturwissenschaftlichen Fakultät.

Kok, K. & Priemer, B. (2023). Using measurement uncertainties to detect incomplete assumptions about theory in an experiment with rolling marbles. *Physics Education, 58(3),* 035007. https://doi.org/10.1088/1361-6552/acb87b

Kok, K., Priemer, B., Musold, W. & Masnick, A. (2019). Students' conclusions from measurement data: the more decimal places, the better? *Physical Review. Physics Education Research, 15*(1), 1-6. https://doi.org/10.1103/PhysRevPhysEducRes.15.010103

Konferenz der kantonalen Erziehungsdirektoreninne und Direktoren (EDK). (1994) *Rahmenlehrplan für die Maturitätsschulen vom 9. Juni 1994. Empfehlung an die Kantone gemäs s Art. 3 des Schulkonkordat vom 29. Oktober 1970.* https://www.edk.ch/de/bildungssystem/beschreibung/lehrplaene

Lippmann, R. (2003). *Students' understanding of measurement and uncertainty in the physics laboratory: social construction, underling concept, and quantitative analysis* [Dissertation, University of Maryland]. ProQuest Dissertation Publishing.

Loidl, H. (2021). Entwicklung und Evaluation von Unterrichtseinheiten zum Thema Messunsicherheiten. [Masterarbeit, Universität Wien]. Universität Wien.

Masnick, & Morris, B. J. (2008). Investigating the development of data evaluation: the role of data characteristics. *Child Development, 79(4),* 1032–1048. https://doi.org/10.1111/j.1467-8624.2008.01174.x

Mayring, P. (2000). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (7th ed.). Weinheim: Deutscher Studien Verlag.

Mayring, P. (2016). *Einführung in die qualitative Sozialforschung: Eine Anleitung zu qualitativem Denken* (6th ed.). Weinheim Basel: Beltz.

McKenney, S. & Reeves, T. C. (2018). *Conducting educational resign Research (Second edition.).* Boca Raton, FL: Routledge.

Millstone, E. & Van Zwanenberg, P. (2000). A crisis of trust: for science, scientists or for institutions? *Nature Medicine, 6*(12), 1307–1308. https://doi.org/10.1038/82102

Ministry for Education – Rheinland Pfalz. (2022). *Curriculum for upper Secondary.* https://naturwissenschaften.bildung-rp.de/fileadmin/user_upload/naturwissenschaften.bildung-rp.de/Lehrplaene/Physik_SekII.pdf

Ministry for Education, Youth and Sport - Brandenburg. (2021). Curriculum for upper Secondary. https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/rahmenlehrplaene/gymnasiale_oberstufe/Materialien_RLP_GOST_Nawi/2021_12_01_RLP_GOST_Teil_C_Physik.pdf

Möhrke, P. (2020). Messunsicherheiten im Physikunterricht. Befragung von Lehrkräften in Baden-Würtember. In S. Habig (Ed.), *Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Wien 2019* (876-879) Essen: Universität Duisburg.

Nagel, C. (2017, 17. November). *Auswertung und Dokumentation experimenteller Daten. Der Umgang mit Messunsicherheiten von der Planung eines Experiments bis zur Publikation analysierter Messdaten*. [Skriptum zur gleichnamigen Vorlesung des Autors aus dem Sommersemester 2016 – _Universität Wien]. https://moodle.univie.ac.at/pluginfile.php/3792555/mod_resource/content/6/Auswertung- Skript.pdf

Nagel, C. (2021). Sicher ist sicher! Fachliche Klärung für die didaktische Rekonstruktion von Messunsicherheiten im Unterricht. *Plus Lucis, 1*(4), 7-11.

Nagel, C. (2023). Trustworthiness as central design principle for introducing uncertainties of measurements to students. In M. Kireš, et al. (Eds.), *Physics learning promoting culture and addressing societal issues:* GIREP-EPEC Conference 2023. [paper submitted]

Nagel, C., Lux, B. & Steindl S. (2021). Die Thematisierung von Messunsicherheiten im Physikunterricht - Eine Umfrage. *Plus Lucis, 1*(4), 4-6.

National Academies of Sciences. 2012. *A framework for K-12 science education: practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

Nedden, M. & Priemer, B. (2020). Aus der physikalischen Forschung in die Schule: Verfahren zur Beschreibung von Unsicherheiten und zur Vermeidung von Bestätigungsfehlern, *Naturwissenschaften im Unterricht - Physik 177/178*, 23-27.

Neumann, S. (2021). Bin ich wirklich schneller als mein Sitznachbar? *Plus Lucis, 1*(4), 36-37. https://www.pluslucis.org/Zeitschrift.html

Oser, F. & Baeriswyl. F. (2001). Choreographies of teaching: bridging instruction to learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed. 1031-1065) Washington: American Educational Research Association.

Petts, P., Swift, G., Peter & Nightingale, S. C. (2021). Evaluation of student understanding of uncertainty in level 1 undergraduate physics laboratories. In *10th International Conference New Perspectives in Science Education*, 413-419. https://conference.pixel-online.net/NPSE/files/npse/ed0010/FP/7220-STEM5059-FP-NPSE10.pdf

Pols, F., Dekkers, P., & de Vries, M. (2019). Introducing argumentation in inquiry – a combination of five exemplary activities. *Physics Education, 54*(5), 055014. doi: 10.1088/1361-6552/ab2ae5

Popper, K. (1934). *Logik der Forschung, Nachdruck der 10. Auflage (2002)*. Tübingen: Mohr Siebeck Verlag.

Priemer, B. & Hellwig, J. (2018). Learning about measurement uncertainties in secondary education: a model of the subject matter. *International Journal of Science and Mathematics Education, 16(1),* 45-68. https://doi.org/10.1007/s10763-016-9768-0

Priemer. (2022). *Unsicherheiten, Aber Sicher! Vom kompetenten Umgang mit ungenauen Daten.* Berlin: Springer. https://doi.org/10.1007/978-3-662-63990-0

Reinmann, G. (2018). *Reader zu Design-Based Research (DBR).* Universität Hamburg. Hamburger Zentrum für Universitäres Lehren und Lernen (HUL). https://gabi-reinmann.de/wp-content/uploads/2018/06/Reader_DBR_Juni2018.pdf

Rowland, J., Estevens, J., Krzewińska, A., Warwas, I., & Delicado, A. (2022). Trust and mistrust in sources of scientific information on climate change and vaccines. *Science & Education, 31*(5), 1399–1424. https://doi.org/10.1007/s11191-021-00304-0

Ryder, J. & Leach, J. (1999). University science students' experiences of investigative project work and their images of science. *International Journal of Science Education, 21*(9), 945-956.

Schecker, H., Hopf, M. & Wilhelm, T. (Eds.). (2021). *Unterrichtskonzeptionen für den Physikunterricht. Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis.* Berlin: Springer. https://doi.org/10.1007/978-3-662-63053-2

Schecker, H., Wilhelm, T., Hopf, M. & Duit, R. (Eds.). (2018). *Schülervorstellungen und Physikunterricht. Ein Lehrbuch für Studium, Referendariat und Unterrichtspraxis.* Berlin: Springer. https://doi.org/10.1007/978-3-662-57270-2

Stake, R. (2010). *The art of case study research (15th ed.).* Thousand Oaks, Calif: Sage Publications Inc.

Strategic Visioning Expert Group, (2020). *PISA 2024 - strategic vision and direction for science.* OECD-Publications. https://www.oecd.org/pisa/publications/PISA-2024-Science-Strategic-Vision-Proposal.pdf

Tesch, M., Duit, R. (2002). Zur Rolle des Experiments im Physikanfangsunterricht. In V. Nordmaier (Ed.), *Didaktik der Physik: Beiträge zur Frühjahrestagung der DPG.* Berlin: Lehmanns Fachbuchhandlung.

Van der Bles, A., Van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences - PNAS, 117*(14), 7672–7683. https://doi.org/10.1073/pnas.1913678117

Van der Linden, S. & Ragnar E. Löfstedt. (Eds.). (2019). *Risk and Uncertainty in a post-truth society.* Routledge. https://doi.org/10.4324/9780429280290

Wagner, S., Maut, C. & Priemer, B. (2021). Thermal expansion of water in the science lab - advantages and disadvantages of different experimental setups. *Physics Education, 56*(3), 035022. https://doi.org/10.1088/1361-6552/abeac4

Wiesner, H., Schecker, H. & Hopf, M. (2011). *Physikdidaktik kompakt* (4th ed.). Aulis-Verlag.

Appendix

Find the curriculum and all teaching materials at:
https://aeccp.univie.ac.at/lehrer-innen/