Contributing factors to the improvement of conceptual understanding in a computer-based intervention in Newtonian dynamics

Jannis Weber[*] and Thomas Wilhelm[†]

Department of Physics Education Research, Goethe University Frankfurt, Max-von-Laue-Straße 1, Frankfurt/M, Hesse, Germany

(Received 3 October 2023; accepted 19 August 2024; published 15 October 2024)

Students experience many difficulties learning the fundamental relationships in Newtonian mechanics, partly due to preexisting mental models that originate from their everyday lives. These preconceptions often persist even after instruction in mechanics and lead to a supposed incompatibility between physics lessons in school and personal experiences. This article presents research that focuses on students' concept knowledge and conceptual understanding in the field of Newtonian dynamics. Interventions are presented that discuss authentic experiments with friction using different software as a tool. Overall, the interventions lead to an increase in conceptual understanding with a large effect size. A hierarchical linear model (HLM) is used to identify factors that lead to a larger or smaller increase in conceptual understanding of the topic. With the HLM, we infer that the pretest score, the students' interest in how physical properties interrelate (interest in "theoretical physics"), prior physics performance (in kinematics in this case, measured in the physics grade), and the (intrinsic and extraneous) cognitive load influence the post-test score. We discuss possible explanations as to why these factors influence students' conceptual change.

DOI: 10.1103/PhysRevPhysEducRes.20.020130

I. INTRODUCTION

Newton's theory of motion is one of the most fundamental theories in physics and forms the basis for many other theories. It is, therefore, of great importance to physics students of all ages [1]. A lot of time is spent in schools teaching Newton's laws of motion, but studies show that students' conceptual understanding is often insufficient after the relevant lessons [2]. Students' misconceptions of mechanics are particularly persistent and well researched [3]. One reason could be that mechanics is closely related to students' everyday experiences and therefore the misconceptions can become especially entrenched over the course of their lives.

Physics teachers are faced with the task of teaching students physically correct and buildable concepts, reinforcing these concepts, while reducing unwanted ones. Although students will not entirely lose their naïve preconceptions [4] (nor should they for everyday communication), according to Posner *et al.* they must be dissatisfied with them and see the physical model presented as more intelligible, plausible, and fruitful [5]. Experience has shown that this is difficult.

Friction is the key to recognize Newton's laws in everyday life. An object, that moves with a constant velocity at a certain time will always come to rest because there is a force acting on it (air resistance, sliding friction, etc.), not because it is an intrinsic property of the object. For example, most students assume that a force must act in the direction of motion for something to move [6]. This is natural from an everyday perspective, as on a bicycle you must constantly pedal to maintain a certain velocity. Without considering friction in these situations, the students see no need to switch from their preconceived concepts to the desired Newtonian concepts, as their naïve mental models give the correct result and explain the situation sufficiently well. However, the discussion of frictional forces poses mathematical problems in physics class, especially in high school and below. Authentic motions, in which several forces, possibly dependent on speed or location, may act on the object, cannot be investigated quantitatively without aiding tools. The computer can be such a tool [7]. Overall, digital tools are proven to have a positive effect on learning gain in the natural sciences and mathematics [8]—especially when used collaboratively in small groups [9]. They should be used additionally and not instead of conventional methods. Effect sizes are the highest in physics and in grades 11–13 [10].

Because of the requirements mentioned above, we focus on two specific software tools, which are especially useful in Newtonian mechanics. On one hand, the computer can be used to model motions (computational modeling,

[*]Contact author: weber@physik.uni-frankfurt.de
[†]Contact author: wilhelm@physik.uni-frankfurt.de

CM) [11]. Here, the user implements the forces acting on a body and creates the mathematical model associated with a particular motion, using the standard model of Newtonian mechanics [12]. The software numerically solves the arising differential equations (computational physics). To overcome the above-mentioned challenges, friction should always be considered. This is possible since the software can help by numerically solving the arising differential equations, which are often not possible to solve by hand. Whether friction needs to be included in the final model or not can be decided by the users themselves. While using CM, first a theoretical mathematical model of the physical reality is created and then implemented. It can then be compared with real data and be iteratively improved in the process until the user considers it sufficiently good [13]. Research indicates a positive effect of using computational modeling in different areas depending on the age of the students, the software used, and the goal of the intervention (see Sec. II B).

The computer can also be used to measure and evaluate experiments in different ways. The measurement data can be displayed and interpreted, with a qualitative discussion of the influence of friction. An example of this is video motion analysis (VMA), which enables an easy-to-implement analysis of everyday motions [14]. In this case, measurement data from experiments are explained with the underlying theory, focusing on friction to illustrate that Newton's laws apply to everyday life. No mathematical knowledge is required to use VMA this way. Using VMA was, among other effects, proven to reduce cognitive load thus increasing the learning gain in kinematics and is associated with a higher motivation in the taught subject (see Sec. II C).

We hypothesize that an intervention that takes research on students' preconceptions into account and uses the computer to analyze or model authentic motions with different frictional forces (without requiring students to have in-depth knowledge of differential equations and other advanced mathematics) will be successful in reinforcing the correct physical theory. This should lead to an increase in conceptual understanding that is higher than in traditional approaches. However, several factors can influence the learning gain in the process. In the following chapters, research is presented that aims to determine whether such an intervention is indeed successful, and which factors influence the success of the intervention with $N = 274$ 11th-grade students from German high schools.

II. THEORETICAL BACKGROUND

A. Conceptual change

The process of transitioning from preexisting conceptions to physically viable ones is called conceptual change. There is plenty of research on the conditions in which conceptual change can be successful [15]. No general consensus exists, partly because of different frameworks of the process of learning itself. One of the most frequently referenced models of conceptual change is from Posner *et al.* [5]. The authors name the following conditions for successful conceptual change: The student must be dissatisfied with the prior conception and the new conception must be intelligible, plausible, and fruitful [5].

According to different studies [16,17], even experts in their respective fields never lose their naïve preconceptions. Instead, they learn to inhibit those in favor of the correct conception. Therefore, Potvin *et al.* [17] prefer the term "conceptual prevalence" over "conceptual change." The goal of successful physics lessons should then be to weaken misconceptions and strengthen the desired ones. The most successful strategy depends on the conceptual change framework. While some authors suggest, that starting with connectable parts of the preconception and gradually transferring to the wanted concept is optimal [18], others suggest that a cognitive conflict is the most successful strategy, since students must view their preexisting mental model as insufficient or wrong [19]. Both strategies do have advantages and disadvantages. Manifold research and the conceptual prevalence framework in particular suggest that weakening the preconception through cognitive conflict before presenting an alternative is counterproductive [16]. In this study, cognitive conflicts were used well after students were presented with the correct laws of motion in school. The cognitive conflicts are generated and dissolved by the interaction with the software itself, not the teacher, which has several advantages, since the software is neutral, and students may be more inclined to express their beliefs while working in pairs of two on a laptop over a discussion in the plenum. It is hypothesized that this leads to a successful transition from unwanted preconceptions to the correct physical theory of motion and therefore to an increase in conceptual understanding in the field of Newtonian dynamics.

B. Other perspectives on conceptual change

Besides the strictly cognitive perspective (Sec. II A), there are broader (sociocognitive) and different (sociocultural) perspectives on conceptual change. While these perspectives were originally separate, some authors combine these into an integrated constructivism [20] for a holistic approach to explain the complex process of conceptual change [21,22].

Within a constructivist and primarily cognitive approach to conceptual change, only the cognitive processes of the individual are focused. This approach does not directly explain other influences on conceptual change. Therefore, suggestions were made to steer toward a "warm" cognitive-affective model of conceptual change [23] (sociocognitive approach) including motivational factors (such as interest in a certain subject, epistemic beliefs, or the subject-specific self-concept or self-efficacy) away from a "cold" and overly rational one [24].

A different approach to conceptual change is the sociocultural one [25], which understands learning as a process of enculturation into a community [26]. Accordingly, the acquisition of knowledge is not explained by the cognitive processes of the individual, it rather is a process linked to a certain context [26] and learning is a process of becoming a member of a particular community [26,27]. Successful learning needs different contexts and discourse practices [28] and knowledge does not transfer between those contexts [29]. Therefore, in this perspective, learning environments should enable social learning and the ability to deal with authentic problems.

To combine the mentioned approaches, Vosniadou proposes to choose the individual's activity in its environment (both physical and social) mediated by symbolic structures as a unit of analysis for conceptual change processes [26]. This study follows a more sociocognitive approach, where the cognitive processes of the individual are focused, which are influenced by social (affective and motivational) factors.

C. Computational modeling

Computational modeling (CM) refers to the "construction of a network of physical concepts and relationships that can describe and predict the behavior of a physical system" [30]. Teachers have been using computer-based modeling to teach physics since the 1980s. The software has changed fundamentally since then [11]. In the early years, it was almost exclusively the teachers who operated the software. They mostly used programming languages such as "Basic" or "Pascal." CM is also possible with spreadsheet programs such as "Microsoft Excel," in both cases using the so-called "method of small steps" ("Euler method"). However, graphical modeling programs have the advantage that the underlying structure of the mathematical model is visually represented. There are also equation-based CM programs where only the acting forces, the initial conditions, and the mass of an object must be entered, while the standard model of mechanics is already implemented, which simplifies the use. Each of these types of computational modeling software has its own advantages and disadvantages and they should be used depending on the goals of the lesson. They all allow students to focus on the physical relationships between the relevant quantities while discussing complex and authentic motions involving frictional forces, without the need to focus on complex numerical solutions.

Research on CM shows that the sometimes far-reaching expectations have not always been met in the past [31,32]. However, it was shown that the use of graphical modeling programs leads to an improvement in the semiquantitative description of mechanical motions compared to normal teaching [33]. Benacka was able to demonstrate a motivating effect of CM with "Excel" on high school students [34–36]. In the university sector, "VPython" [37], a software based on the "Python" programming language, with which simple 3D animations for model building can be created, has become somewhat established and successful [38]. In schools, the use of "VPython" seems to be more problematic [39]. Especially if the learners are to use the software autonomously, an intuitive and easy-to-use interface is important. Therefore, we used the equation-based CM program "Newton-II[1]" here, for which there are no empirical studies so far. In "Newton-II", the focus is on the relationship between the acting forces and the subsequent motion of an object, while the kinematic relationships (between acceleration, velocity, and location) are calculated automatically by the program and are not a central aspect of the intervention.

D. Video motion analysis

Video motion analysis (VMA) is a measurement method that enables easy analyses of two-dimensional motions and the presentation of the acquired data in various ways. Compared to data acquisition systems, the easy-to-use display options are an advantage of VMA. It has been used in physics education for many years [40] and has become increasingly established in the recent past, also due to improved software [41,42]. VMA can be performed on laptops and tablets with different advantages. Overall, studies on VMA have shown that students are more successful in interpreting motion diagrams [43] and in understanding the concept of braking [44] when using it. In addition, positive effects on motivation and curiosity have been measured [45]. It has been shown that the understanding of concepts, especially in challenging tasks in kinematics, is improved more using VMA than through normal teaching, which is explained by a reduction in cognitive load [46]. The use of multiple representations [47], such as graphs or vector arrows in combination with videos, can have a positive effect on learning progress [48]. Overall, however, research on VMA is largely limited to the field of kinematics. In this study, the use of the software "measure dynamics[2]" for laptops in the field of dynamics is investigated.

E. Cognitive load heory

Cognitive load theory (CLT) [49,50] postulates a limited capacity of the working memory, which is important for learning processes. In addition, the time available for processing information is also limited. The working

[1]"Newton-II" is an equation-based CM program. It was created by Dr. Stephan Lück at the Department for Physics and its Didactics at the University of Würzburg and can be downloaded for free: https://did-apps.physik.uni-wuerzburg.de/Did-Apps-Site/newton-ii_info/.

[2]"measure dynamics" is a VMA program. It was created by Michael Suleder and a demo version can be downloaded at PHYWE: https://www.phywe.de/sensoren-software/mess-software-apps/software-measure-dynamics-einzellizenz_2259_3190/.

memory is important for building new schemata and should therefore not be overloaded. This should be considered when designing learning environments. Three types of cognitive load are distinguished [51]. Intrinsic cognitive load (ICL) depends on the complexity of the information to be processed and is thus dependent on the learning task and the learner's prior knowledge. It cannot be influenced by the learning environment. Extraneous cognitive load (ECL) is largely influenced by the design of the learning environment and should be minimized according to CLT, as it is not conducive to the learning process. Germane cognitive load (GCL) is the cognitive load relevant to the learning process, which is responsible for the construction and automation of new schemata and should be as high as possible. CLT is a theory that can provide an explanation of which learning environments are conducive to learning and which lead to unnecessary cognitive load that hinders the building of new schemata. In the context of video motion analysis, a reduction in ECL has been shown to be associated with higher learning gains [52].

III. RESEARCH QUESTIONS

This study is conducted with students who have already covered Newton's laws of motion in their school lessons. The goal is to investigate to what extent a subsequent intervention on Newtonian dynamics with computational modeling or video motion analysis is effective for deepening and improving their conceptual understanding.

1. Is an intervention based on the known preconceptions using a computer to discuss authentic motions including friction successful in improving the conceptual understanding of Newtonian dynamics after the introduction in school?

We hypothesize, that this is the case. When an appropriate intervention achieves the expected learning gains, the central question discussed in this article is, which specific factors lead to a higher or lower learning gain and how that fits into the sociocognitive perspective on conceptual change.

2. Which factors can be identified that influence the students' success in the intervention and how do those factors affect their conceptual change?

From a sociocognitive standpoint, different affective variables could impact the way knowledge is constructed. As the interventions use a more theory-based approach (computational modeling) and a more experimental approach (video motion analysis), it is hypothesized that learners who are more interested in the interrelationships between physical quantities will achieve greater learning gains with computational modeling, while learners who are more interested in experimentation will learn more with video motion analysis. Recommendations for the use of the two methods in the classroom could be derived from the results. Furthermore, affective variables, such as interest in physics and computer affinity, may also have an impact on conceptual change, though it is unclear to what extent. The grades in mathematics and physics might influence learning gain, as well as the physics-related self-concept.

Additionally, the cognitive load should theoretically affect the learning outcome of the intervention as well.

IV. DESIGN

To answer the research questions, a quasiexperimental study using a pre-post design was conducted. Two separate interventions were created, one with a more theoretical approach using computational modeling to model different motions and compare the model to experimental data and the other one using video motion analysis to analyze the motion and derive the acting forces from the data. Both focus on the fundamental relationship $\vec{a} = \frac{\sum \vec{F}}{m}$, each method approaches the equation from a different direction. The two interventions were otherwise similar (Sec. IV B). They took place in the physics laboratory of the Goethe University in Frankfurt/M. A test was conducted in the last lesson in school before the visit and one immediately after the intervention to measure, among other variables, the change in conceptual understanding that can be ascribed to the intervention.

A. Test instrument

For this study, a test was created and piloted. It is in German language and consists of several different parts.[3] The subscales reported in the following chapters were found using an exploratory (and a subsequent confirmatory) factor analysis[4] [54]. The model fit was evaluated and a test for tau-equivalency was conducted [55]. In cases where tau equivalency had to be rejected in unidimensional scales, the composite reliability (congeneric reliability, [56]) was calculated, else Cronbach's alpha was used [57]. The only multidimensional variable, included in the hierarchical linear model, was the conceptual understanding of Newtonian dynamics. Here, a bifactor model was used, since the items measure the overall understanding of Newton's first two laws and consist of three subscales (Newton's first law; Newton's second law, force to motion; Newton's second law, motion to force).

1. Conceptual understanding

The main part of the test instrument consists of testing conceptual understanding of Newtonian dynamics.[5] The FCI

[3]The complete test instrument for the overall study includes some additional items, which are irrelevant for the results presented in this study. For a discussion of those items and the complete test, see Ref. [53].

[4]Principal component analysis with varimax rotation and Kaiser normalization (see Ref. [53] for a more in-depth-explanation in German).

[5]See also Ref. [11].

TABLE I. Newtonian dynamics.

Scale name	Number of items	Pretest	Post-test	ω_t
Conceptual understanding of Newton's first two laws	15	✓	✓	0.84

(Force Concept Inventory) [58] is the best-known test for conceptual understanding of forces. It reliably tests force understanding [59] and offers high comparability to other studies [60]. However, the items in the FCI cannot be reliably assigned to different content areas that would allow for a more differentiated view [2,61,62],[6] which was necessary for the overall study, the test was created for.[7] Since the FCI contains fitting items, it served as the basis for creating the test used here. Of the 15 items of the test, 9 are from the FCI. Of these, three were adopted in their original form, four were modified, and two were supplemented by a second level, in which the reasoning for the response in level 1 had to be selected. Three items were newly created, two are from Wilhelm [64] in accordance with Flores *et al.* [65], and one is taken from Warren [66] in a modified form. All items are evaluated dichotomously, and a two-level item is considered correct only if the answer and the reasoning are correct. After piloting with $N = 85$ students, we identified three content areas using a scree test and principal component analysis that contribute to the overall understanding of Newton's first two laws (Table I).

The overall test instrument also included four items on acceleration graphs ($\alpha_C = 0.91$) from the FMCE (Force and Motion Conceptual Evaluation) [67] in the version of Wilhelm [47,68], to measure the conceptual understanding before the intervention. An improvement here was not the goal of the study. In addition, the content area "Newton's third law" with four items was added as a control scale from the FCI in both pre- and post-test, which already shows a high internal consistency in the FCI [62,69] and was therefore not part of the pilot study. Newton's third law is not part of the intervention. No change in conceptual understanding is expected here. A significant increase in the items in the scale "Newton's third law" could indicate problematic effects regarding the test. Students could have improved because they knew the items in the post-test, were more focused, etc. We added those items to rule out this possibility.

[6]Research also indicates that it makes a difference (in difficulty) to students whether they infer the acceleration from the forces or vice versa [63].

[7]The different content areas are not relevant for the results presented in this paper but were the reason for a creation of a new test instrument.

2. Rating of the intervention

To be able to test affective variables, we adopted items from Laukenmann *et al.* [70]. The scale "relevance of the topic" contains items that ask about the importance of what was learned and students' desire to learn more, while the "fun of the intervention" scale is about the experienced enjoyment of the day. In addition, we created our own items to evaluate the used software. In these items, the user-friendliness of the software, the suitability for learning physics, and the desire to use it in school lessons are queried. Students had to answer the items on a five-point Likert scale. The pilot study here also took place with $N = 85$ students (see Table II).

3. Cognitive load

To measure cognitive load, an established test instrument was adopted from Becker *et al.* [52], who created and evaluated a German version of the English CLS by Leppink *et al.* [71]. The items were adapted to our context.[8] Even though the CLS has been shown to be robust in similar cases [72,73], we conducted a factor analysis to confirm the scales with similar results as [72] and calculated the Cronbach's alphas of the three scales (Table III). Sufficient validity and reliability can therefore be assumed. The items are answered on a six-point Likert scale.

4. Other affective variables

Various additional variables were measured, which are hypothesized to influence conceptual change. These include the mathematics and physics grade, subject-specific self-concept [76], subject interest [77], interest in interrelationships between physical quantities ("theoretical physics," own items), interest in experiments in physics ("experimental physics," own items), and computer affinity [78] (see Table IV). The students had to answer all these items on a five-point Likert scale.

In total, the pretest consisted of 50 items and the post-test of 52 items.[9] Students were able to complete the questionnaire in about 30–35 min.

B. Interventions

The design of the interventions was based on findings in the area of digital media use in the classroom [7,9,10,79]. The interventions lasted 3.5 h and students participated after their lessons in Newtonian mechanics in school (Fig. 1). They received a workbook, which led them through the exercise with different tasks and questions. In both interventions, four

[8]We only changed "gleichförmige Bewegung" (= uniform motion) to "Dynamik" (= dynamics).

[9]There were some additional items on model understanding in physics that are not relevant for the results discussed in this article.

TABLE II. Different affective scales regarding the intervention.

Scale name	Number of items	Pretest	Post-test	Reliability[a]
Usability of the software	5	X	✓	0.78[b]
Relevance of the subject	3	X	✓	0.78[c]
Enjoyment of the intervention	4	X	✓	0.85[c]

[a]Where tau-equivalency had to be rejected, reliability was calculated using the composite reliability. Else, Cronbach's alpha was used.
[b]Composite reliability.
[c]Cronbach's alpha.

TABLE III. Three different subscales of the cognitive load.

Scale name	Number of items	Pretest	Post-test	Reliability[a]
Intrinsic cognitive load	3	X	✓	0.78[b]
Extraneous cognitive load	2(3)[d]	X	✓	0.72[c] (0.59)[d]
Germane cognitive load	4	X	✓	0.90[b]

[a]The reported reliability coefficients are from our own data. It should be noted, that while the reliability of the ICL and GCL scales is good, the reported Cronbach's alpha of the ECL scale from the original questionnaire with ten items is problematic, which is a common problem in different versions of the CLS [74]. A factor analysis showed that, in our case, one of the three items does not load on a certain factor. We removed that item (see also Ref. [72] for a similar approach). The ECL scale consisting of the remaining two items has a reliability coefficient of 0.72 (Spearman-Brown ρ_{y1y2} [75]) and was used in the HLM.
[b]Composite reliability was used since tau-equivalency had to be rejected.
[c]Spearman-Brown.
[d]Values from the original questionnaire from Becker et al. in brackets.

TABLE IV. Additional affective variables.

Scale name	Number of items	Pretest	Post-test	Reliability[a]
Subject-specific self-concept	4	✓	X	0.93[b]
Interest in physics lessons	4	✓	X	0.87[b]
Interest in theoretical physics	5	✓	X	0.82[c]
Interest in experimental physics	3	✓	X	0.82[c]
Computer affinity	4	✓	X	0.90[c]

[a]Where tau equivalency had to be rejected, reliability was calculated using the composite reliability. Else, Cronbach's alpha was used.
[b]Cronbach's alpha.
[c]Composite reliability.

experiments were conducted. Each experiment and the associated tasks took 45–60 min to complete. The students did not know the software they had to use in the intervention. Therefore, in the first of the four experiments, the teacher introduced how to use it.

In the CM cohort, following the experiment in the plenum, students developed hypotheses about which forces might have acted on the object during its motion. These hypotheses were explicated and tested using the software "Newton-II" in pairs. After an initial test of the model, the students could adjust their hypotheses by changing the implemented forces. Afterward, a comparison of the model with real kinematic data took place, which was inserted into the software. Usually, the model differed from the real data at first, either due to errors in the model or because friction was not yet accounted for. Students were left with the choice if they want to or should include frictional forces. While the Styrofoam ball (Sec. IV B 3), for example, is heavily influenced by air friction a model without friction fits the trajectory of the steel ball sufficiently well. The trajectory of the cone (Sec. IV B 1) and the trolley (Sec. IV B 2) can only be explained with friction. The visible difference between the model without friction and the data shows that friction plays an important role and that an object does not slow down without a force acting on it. This is important to strengthen Newtonian and to weaken

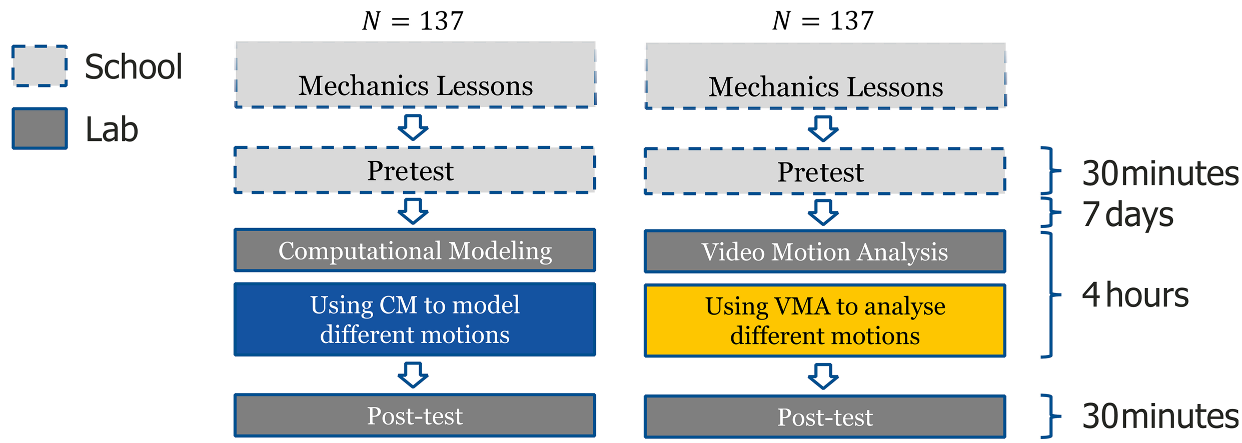

FIG. 1. Structure of the study. The assignment to one intervention is made by whole classes.

Aristotelian or impetus views. After adjusting the model again, the students used the model to answer questions concerning the relationships between the relevant quantities to consolidate the gained knowledge. The whole process was guided by the workbook. The goal was to strengthen Newtonian argumentation,[10] deepen the structural understanding of Newtonian dynamics,[11] and increase the conceptual understanding of Newton's first two laws.

After conducting the initial experiment, the VMA group started with a video clip of the experiment, from which they obtained kinematic data with the help of an automatic video motion analysis with the software measure dynamics.

These data were processed and displayed in different ways, depending on the task in the workbook (velocity and acceleration arrows in the video, graphs created synchronously with the video, and stroboscopic images). The interpretation of the data illustrated the relationship between the kinematic quantities and allowed to infer the total force acting from the acceleration. Students then had to discuss the components of the total force. Frictional forces are discussed here as well, though only qualitatively. It is emphasized that the observed motion can only be explained by other forces acting on the object and discussed how the trajectory would look without friction. The acceleration in the opposite direction of motion of the trolley, for example, can only be explained by friction since there are no other forces acting (in the axis of motion). Analogous to the computational modeling group, the goal was to increase Newtonian argumentation by focusing on the connection between acceleration and forces.

Both groups conduct experiments and both carry out theoretical considerations. The fundamental difference between the two groups, besides the software used, is the direction of reasoning. One group approached the fundamental relationship between forces and acceleration from theory and modeled reality (i.e., the modeling of forces and the consideration of their effects on acceleration, velocity, and location), and the other group started from the measured data (location), analyzed it (the software calculated velocity and acceleration), and derived the forces from there. Therefore, the CM approach tends to match more to the classical approach of theoretical physics (modeling reality based on a theoretical model and evaluating that model with real data), while the VMA approach corresponds more to experimental physics (evaluation of experimental data to derive theoretical insights), even if both groups intensively use the computer. The general approach was identical for all experiments performed in the intervention.

1. Falling cone

The initial experiment was a falling cone [81,82]. The relevant forces were a constant gravitational force acting downward and an increasing ($\sim v^2$) air friction force acting upward. The falling cone moves at a constant speed at the end of the motion since the air friction force increases until the sum of all forces acting on the cone is zero. With this experiment, the students also learned to use the respective program by observing the teacher and copying the necessary steps. After conducting the experiment in the classroom, all students had to describe the motion first.

In the CM cohort, the software is then used to create a model. Usually, gravity was the first suggestion by students and was implemented first (F_G). The initial conditions were given by the workbook. Then, a first evaluation of the model took place. If no obvious errors are visible in the model (e.g., the cone flies upward or not at all), real experimental data are included in the same diagram to compare the model to reality (Fig. 2). Possible differences between model and reality must be explained (the biggest factor here is air friction). Air friction is included next (F_A) in a simplified form ($k \cdot v^2$). The model is compared to the data again (adjusting k in the process). When students

[10] An increase in physically correct argumentations due to computational modeling has already been observed [80].
[11] This was shown in Ref. [47].

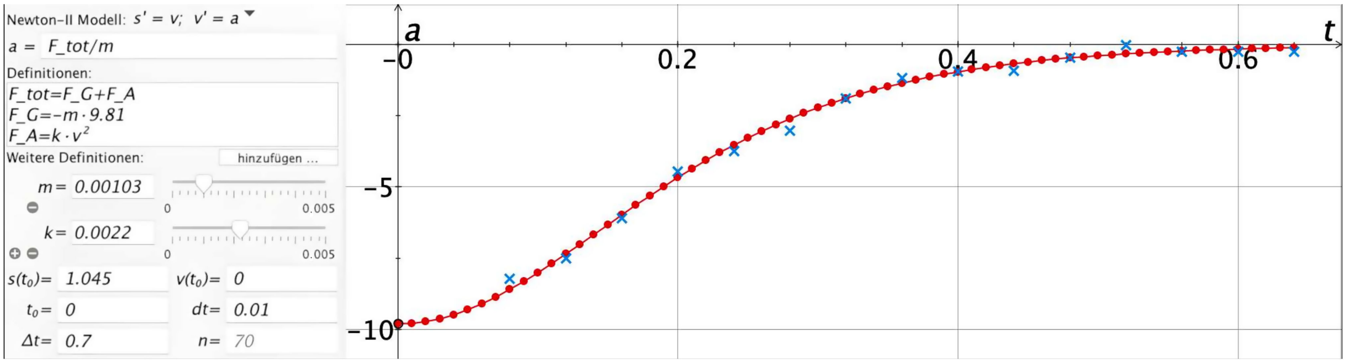

FIG. 2. Screenshot of a computational model (left) of the falling cone and the corresponding acceleration graph (right, red dots) in "Newton-II." The same diagram also shows the data from a real experiment (blue crosses) for comparison.

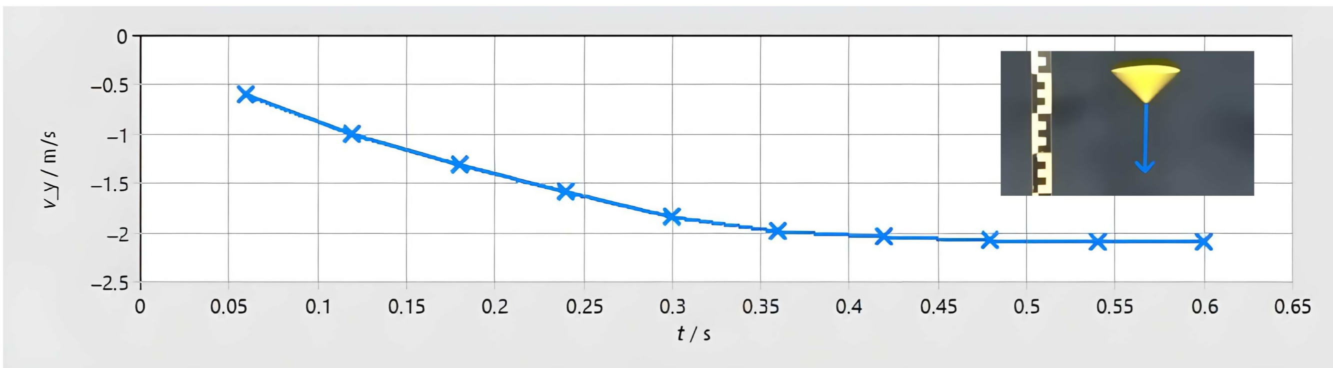

FIG. 3. Screenshot of the measured velocity of the falling cone in "measure dynamics." Additionally, a screenshot from the video with a velocity vector is shown that moves with the cone in the video.

are satisfied with their model, they answer questions about the gravitational force, the air friction, the sum of all forces, and consequently the acceleration,[12] which results from these, optimally by using the model. The whole process is guided by the workbook and in the end, a discussion in the plenum takes place to compare and evaluate the results. In all following experiments, students worked autonomously in pairs.

The beginning is the same in the VMA group. After the experiment, the students were given a video of the experiment to gather data using "measure dynamics." They analyze the vector arrows of the velocity and acceleration and the respective graphs (Fig. 3) during the fall to conclude that the absolute value of the acceleration decreases in the process to (nearly) zero. Then, they must explain that observation with the acting forces on the cone. In the end, they answer questions about the acceleration and conclude that the (absolute value of the) sum of all acting forces must be decreasing, which can only be explained by increasing air friction since the gravitational force is constant. All following experiments are carried out in an analogous way.

2. Trolley

After the introduction, the students independently worked with the program in pairs of two after the experiment was conducted in the plenum. The second experiment was a trolley traveling on a straight track in the horizontal plane that is initially accelerated by a weight connected to the trolley by a rope. About halfway through the motion, the weight comes to rest on the ground, leaving only frictional forces acting on the trolley. Without a force in the direction of motion, the trolley continues to move. Frictional forces slightly accelerate it. Especially the second phase of the motion is important to counter the misconception that a force in the direction of motion is necessary.

3. Projectile motion

The first two-dimensional motion was a projectile motion created by a catapult. It is of particular importance

[12]The choice is always between "constantly zero," "constant ($\neq 0$)," and "changing." The idea is, that students recognize the pattern, that they must always tick the same box for acceleration and the sum of all acting forces.

that the flight of a steel ball, where a model without air friction describes reality sufficiently accurately, is compared with the flight of a Styrofoam ball, whose trajectory is strongly influenced by air friction. The specific shape of the trajectory (parabola) was not the primary focus, though, rather the general relationship between acceleration and the sum of the acting forces. Students are usually surprised, that there is no force acting in the horizontal direction (in the case of the steel ball without relevant air friction) when the ball left the catapult. This can easily be seen by looking at the forces in the model that describe the motion well or by analyzing the acceleration vectors when using VMA.

4. Circular motion

The last experiment was a circular motion in which an electromagnet, which was connected to an engine, held a steel ball. The steel ball is accelerated in a circle with constant speed. After a certain time, the electromagnet is switched off—the ball continues to roll straight ahead. Typical student ideas here are that the centrifugal force pushes the ball outward or that there is no acceleration during the uniform circular motion. The students had to predict the trajectory after switching off the electromagnet before trying it out. The straight-line rolling with (almost) constant speed had to be explained in the course of the work phase with the absence of any forces.

C. Sample

Originally, 630 or so German high school students (11th grade) from 33 school classes were scheduled to take part in the study. Due to school closures in the Covid-19 pandemic and subsequent changes in the schedule, only about half of them could participate. There are $N = 274$ fully complete datasets from 20 different classes, which are analyzed in the following hierarchical linear model.

V. RESULTS

A. Efficacy of the interventions

To investigate whether the described interventions are successful in improving conceptual understanding, the overall score in the items regarding conceptual understanding of Newton's first two laws in the post-test (t_2) is compared to the pretest (t_1) in both groups (15 items, Fig. 4). Items regarding acceleration graphs (four items) and Newton's third law (four items) were not included since an improvement here was not the goal of the interventions. Both groups significantly improve their score ($p < 0.001$) with a large effect size ($d_Z = 1.04$ and $d_Z = 1.07$) [83].

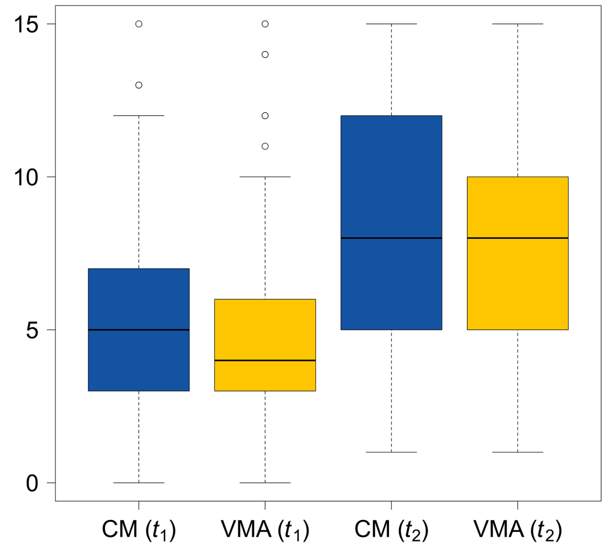

FIG. 4. Boxplot of the score (maximum 15 points) in pretest and post-test in both groups (CM: computational modeling, VMA: video motion analysis). A box plot shows the minimum, the lower quartile, the median, the upper quartile, the maximum, and the outliers (circles).

B. Hierarchical linear model

The participating students are nested in classes. Those classes have different teachers and are from different schools. These factors can have an influence on the outcome, even though the interventions took place in the university, were not conducted by the teacher of the class and students were paired in groups of two, which reduces student-student and student-teacher interaction effects and other influences like class size. Still, considering every student as independent would overestimate the sample size [84] due to their existing codependencies.

One way to take the hierarchical structure into account is a hierarchical linear model (HLM) [85]. Here a two-level hierarchical model is used since only one teacher participated with two classes and therefore the teacher variables are included in the class level. There is no consensus about the statistical requirements of using an HLM [86,87]. Recommendations for using such a model range from 10 to 50 units at the highest level. Newer research suggests that the stability of an HLM is sufficient when using ten units [84,88–90]. Here, there are 21 units on the second and 274 units on the first level.

In the following paragraphs, the procedure of creating the HLM is presented [84,91] and the most important models are shown before discussing the final model, which gives insight into the variables that influence learning gain in conceptual understanding in the presented interventions. To answer the research question, which variables influence the learning gain of the students in the described interventions, a step-up strategy [85] is used to find the best model for the observed data. The following variables were measured, which could potentially influence the learning outcome (Table V).

Different parameters will be used to measure the model fit. The Bayesian information criterion (BIC) measures the power of the model to explain the dependent variable, using

TABLE V. Overview of the possible predictors for the post-test score for Newtonian dynamics.

Level	Name	Scale	Value range
1	Pretest score (Newtonian dynamics)	Interval[a]	[0;15]
1	Interest in physics lessons	Interval (Likert)[b]	[1;5]
1	Subject-specific self-concept	Interval (Likert)[b]	[1;5]
1	Computer affinity	Interval (Likert)[b]	[1;5]
1	Interest in theoretical physics	Interval (Likert)[b]	[1;5]
1	Interest in experimental physics	Interval (Likert)[b]	[1;5]
1	Usability of the software	Interval (Likert)[b]	[1;5]
1	Relevance of the subject	Interval (Likert)[b]	[1;5]
1	Enjoyment of the intervention	Interval (Likert)[b]	[1;5]
1	Gender	Nominal	d, f, m
1	Prior physics grade	Interval[c]	[0;15]
1	Prior math grade	Interval[c]	[0;15]
1	Intrinsic cognitive load	Interval (Likert)[b]	[1;6]
1	Extraneous cognitive load	Interval (Likert)[b]	[1;6]
1	Germane cognitive load	Interval (Likert)[b]	[1;6]
2	Group	Nominal	CM, VMA
2	Test score of the teacher	Interval[d]	[0;15]
2	Average pretest score of the class	Interval[d]	[0;15]

[a]Points in a concept knowledge test are often assumed to have an interval scale, which is only true if the points are equidistant. The classical test theory does not provide such a distribution. Here, a Rasch analysis [92] has been conducted *a posteriori*. The person's ability is linearly linked to the test score except for the edges of the value range (≤ 1 and ≥ 14 points). This implies that the points can be seen as interval scaled with only a minor error.
[b]Likert-Scales are ordinal scales as well. Typically, they are seen as interval scales nonetheless [93]. The statistical error is assumed to be small.
[c]Like points in a concept test, school grades can technically not be seen as interval scaled. As discussed above, the assumption can be made with a sufficiently small statistical error.
[d]Similar considerations as in [a] have to be made when using the test score of the teachers of the participating classes.

the log-likelihood value, and punishes sample size and the number of parameters in the model more than the AIC. The lower the BIC (and AIC), the better the model. There are different ways to estimate these model parameters. The restricted maximum likelihood method is best suited for models with few level 2 elements [94]. To compare different models, a refit using the maximum likelihood method must be used, though [95]. The reported BIC in this article is the latter. A different measure for the goodness of fit of a model is the pseudo-R^2, which can be calculated from the log-likelihood of the respective model in relation to that of the null model [85]. It varies from 0 to 1, where 1 corresponds to a good model fit.

1. M0: Null model

First, the so-called null model is created, which has certain parameters (Table VI). With the null model, it is possible to calculate the variance that can be explained by factors on each level (Table VII). Here, up to 25% ($\mathrm{ICC}_{\mathrm{Level\ 2}} = 0.25$) of the variance can be explained by the class affiliation. Hartig *et al.* [96] recommend using an HLM in this case since the Level-2 ICC is above 10%.

TABLE VI. Description of the parameters of the null model.

Level and equation	Explanation of the parameters
Level 1: $\mathrm{Post}_{ij} = \beta_{0j} + e_{ij}$	• The index i stands for the respective student, j for the class. • Post_{ij} stands for the post-test score from student i in class j. • β_{0j} is the regression constant (average score in class j). • e_{ij} is the random effect on level 1 (in the null model this corresponds to the difference between the post-test score of student i from class j and the average score of class j.
Level 2: $\beta_{0j} = \gamma_{00} + r_{0j}$	• γ_{00} is the average of the post-test scores from every student in every class. • r_{0j} is the random effect on level 2. In the null model that corresponds to the difference between the average post-test score in class j and the overall average γ_{00}.

TABLE VII. Structure and model fit of the null model.

Model type	Equation
HL-Model	Level 1: $\mathrm{Post}_{ij} = \beta_{0j} + e_{ij}$
	Level 2: $\beta_{0j} = \gamma_{00} + r_{0j}$
Mixed model	$\mathrm{Post}_{ij} = \gamma_{00} + r_{0j} + e_{ij}$
Model fit	BIC = 1415.8

According to the step-up strategy, one variable at a time is included in the model based on theoretical considerations. Then we tested, if the model, which includes the additional variables, explains the observed variance significantly better than the prior model. A variable could have an influence on the intercept and/or the slope of the regression. Interactions between the variables are possible as well. Those are only tested if there is a theoretical reason to assume such an interaction exists.

2. Intermediate models

M1: Main effect "Pre-Test". Due to the fact that the interventions were conducted after the lessons of Newton's laws in school and their relatively short duration, it is hypothesized that the pretest score already predicts the post-test score to a relatively high degree. Therefore, the pretest score ("Pre") is included in the hierarchical linear model first.

The model parameters show that model M1, which includes the pretest score ($N = 4$ parameters, BIC = 1322), explains the observed variance significantly better than the null model ($N = 3$ parameters, BIC = 1416; $\chi^2 = 99.2$, d.o.f. = 1, $p < 0.001$). Therefore, the pretest variable remains in the HLM. The pretest score only significantly influences the intercept of the regression lines, not the slope. At this point, the HLM estimates an improvement of $\gamma_{10} = 0.69$ points in the post-test for every additional point in the pretest.

The influence a certain variable has on the post-test score can change if other variables are added to the model. Since the different variables here have different value ranges, the size of the estimate does not imply a larger or smaller influence compared to another variable. In the final model, the effect size and the normalized value of each relevant variable are given.

M2: Main effect "Group". As discussed before, it is hypothesized that the approach (using computational modeling vs video motion analysis) has an influence on the post-test score due to their different ways of approaching the fundamental relationship between forces and acceleration. Therefore, the factor "Group" is added next to the HLM.

No reduction in the BIC compared to the prior model can be observed (M1: $N = 4$ parameters, BIC = 1322; M2: $N = 5$ parameters, BIC = 1328; $\chi^2 = 0.01$, d.o.f. = 1, $p = 0.91$). The HLM rejects the hypothesis that the type of intervention, which was used here, influenced the post-test score. Model M2 does not explain the variance in the data better than model M1. Therefore, model M2 is rejected, and the factor "Group" is removed from the model. It is possible, that a factor does not significantly influence the post-test score by itself but still has a significant interaction with another variable. Since the "Group" factor is still important for the research questions, it is readded in later models to investigate potential interactions with other variables, where an interaction is expected based on theoretical considerations (see M3).

M3: Main effects "Theory" and "Experiment". As discussed before, one intervention follows a more theory-based approach (computational modeling) while the other one uses a more experimental approach (video motion analysis). It is plausible, that the interest in theoretical physics ("Theory") on one hand and the interest in experimental physics ("Experiment") on the other hand influence the learning gain. It would be sensible to assume, that those with a higher interest in theoretical physics learn more when they are in the group of computational modeling while those with a higher interest in experimental physics learn more in the video motion analysis group. An interaction effect between the variables "Theory" and "Experiment" and the "Group" variable would be expected in this case. First, only the factor "Theory" is added to the model.

The new model M3 ($N = 5$ parameters, BIC = 1299) with the factor "Theory" explains significantly more variance than the old model M1 ($N = 4$ parameters, BIC = 1322), which is shown by the direct comparison of the two models ($\chi^2 = 28.4$, d.o.f. = 1, $p < 0.001$). The model, which includes the interaction between "Group" and "Theory," does not explain significantly more variance in the data than M3 and is rejected ($N = 4$, BIC = 1310; d.o.f. = 2, $\chi^2 = 0.001$, $p = 1$).

The interest in theoretical physics positively influences the post-test score even when the pretest score is already accounted for. Every additional point leads to 1.12 additional points in the post-test. This is independent of the kind of intervention, meaning that the more a student is interested in theoretical physics, the more they learn using the experimental or the more theoretical approach. No interaction between the group and the interest score could be observed.

The variable "Experiment" on the other hand does not have an influence on the post-test at all. Neither does the general interest in the school subject physics or the computer affinity. The models including these variables will not be discussed in detail here.

M4: Main effect "Physics grade" and "Math grade". When using computational modeling in physics, a varying degree of math skills is necessary. To understand if knowledge of mathematics is a necessary prerequisite to learn physics using computational modeling, the influence of the math grade on the post-test score is investigated next.

The math grade does not influence the post-test score. There is also no interaction with the "Group" factor, even though the CM group arguably uses more math (but only in the form of writing down equations) during the intervention. The physics grade, however, has a significant influence on the learning gain. Model M4 adds the variable "P_Grade."

Model M4 ($N = 6$ parameters, BIC = 1293) significantly increased the explained variance compared to M3 ($N = 5$ parameters, BIC = 1299), which can be seen in the direct comparison of the two models ($\chi^2 = 12$, d.o.f. = 1, $p < 0.001$). The model estimates that every additional point in the physics grade improves the post-test score by 0.20 points. The physics grade is therefore being kept in the model.

M5: Main effects "rating of the intervention". It should be assumed from a sociocognitive perspective that the different scales that measure the rating of the intervention by the students lead to a higher learning gain in the intervention. The "usability of the software" (M5a), the "enjoyment of the intervention" (M5b), and the felt "relevance of the subject" (M5c) should lead to a higher learning gain due to less difficulties and more engagement in the task. This is not confirmed by the HLM and none of the mentioned effects remain in the model.

M6: Main effect "Cognitive Load". As discussed in the theoretical background, the influence of cognitive load on learning gain has been observed in similar situations. The cognitive load should, therefore, have an influence on the post-test score here as well. This is confirmed by the HLM. First, the ICL is added to the HLM (model M6a), which significantly improves the model ($N = 7$, BIC = 1288; $\chi^2 = 10.3$, d.o.f. = 1, $p < 0.01$). The model, which additionally includes ECL (Model M6b), explains even more variance in the data than M6a ($N = 8$, BIC = 1289; $\chi^2 = 5.0$, d.o.f. = 1, $p < 0.05$). The BIC is not reduced due to the increased parameters of the model.[13] The best model in terms of the predictive power of the post-test score contains the score of the ECL and the ICL. One additional point in the cognitive load Likert scale of the ICL leads to a reduction of 0.37 points and one additional point in the ECL scale leads to a reduction of 0.39 points in the post-test score. Contrary to the assumptions of the CLT, the GCL is not found to be influencing the post-test score in any way.

3. Summary and final model

The final model (Table VIII) includes all variables that were found to be influential for the post-test score: the pretest score, the interest in theoretical physics, the physics grade, and the intrinsic and extraneous cognitive load.

The HLM estimates 0.53 additional points in the post-test score for every additional point in the pretest score (Table IX). Due to the different scales of the variables, the raw influence on the post-test score is not a meaningful measure to decide which variable has the biggest influence.

TABLE VIII. Structure and model fit of the final model.

Model type	Equations
HL model	Level 1:
	$\mathrm{Post}_{ij} = \beta_{0j} + \beta_{1j}\mathrm{Pre} + \beta_{2j}\mathrm{Theory}$
	$+\beta_{3j}P_{\mathrm{Grade}} + \beta_{4j}\mathrm{ICL} + \beta_{5j}\mathrm{ECL} + e_{ij}$
	Level 2:
	$\beta_{0j} = \gamma_{00} + r_{0j}$
	$\beta_{1j} = \gamma_{10}$
	$\beta_{2j} = \gamma_{20}$
	$\beta_{3j} = \gamma_{30}$
	$\beta_{4j} = \gamma_{40}$
	$\beta_{5j} = \gamma_{50}$
Mixed model	$\mathrm{Post}_{ij} = \gamma_{00} + \gamma_{10}\mathrm{Pre} + \gamma_{20}\mathrm{Theory} + \gamma_{30}P_{\mathrm{Grade}}$
	$+\gamma_{40}\mathrm{ICL} + \gamma_{50}\mathrm{ECL} + r_{0j} + e_{ij}$
Model fit	BIC = 1288.8

Table X shows the main effects of all presented models, the variance explained on both levels and different model fit parameters. None of the mentioned variables significantly influenced the slope of the regression lines, only their intercept. We found no interaction effect between any of the variables. All variables not included in the table were not found to be influential for the learning gain.

Gender does explain some of the variance if only the pretest score and the gender are included in the model[14] (male students have a higher learning gain in this case). If the physics grade and the interest in theoretical physics are included in the model, the difference between the two analyzed genders disappears. Male and female students differ in these variables, which in turn better explain the observed variance than the gender itself. The physics-related self-concept does explain some of the observed variance if the physics grade is not included in the model. The physics grade and self-concept in physics are correlated ($r = 0.49$, $p < 0.001$), which is not surprising, with the former better explaining the observed data. The enjoyment of the intervention and the perceived relevance of the topic lead to a higher learning gain, but only if no other variables are included in the model. The conceptual understanding of Newtonian dynamics of the participating teachers themselves, which in 6 of the 17 complete datasets was under the threshold of 80% (Newtonian thinkers[15]) [58], was not found to be influencing the post-test score of the students. The best model is, therefore, an intercept-only model without factors on level 2.

[13] The AIC, which penalizes the number of parameters less than the BIC, is reduced in M6b compared to M6a.

[14] Only the distinction between male and female students was made since there were too few diverse students to statistically analyze that group.

[15] The mentioned threshold is defined for the FCI. Due to the similarities to the test used here, it can also be used in this case.

TABLE IX. Relevant statistical values for the parameters included in Model M6b. SE: Standard error, d.o.f.: degrees of freedom. γ_{00} is the intercept, γ_{10} the pretest score, γ_{20} the interest in theoretical physics, γ_{30} the physics grade, γ_{40} the ICL, and γ_{50} the ECL.

Parameter name	Estimate	SE	d.o.f.	t value	p value	95% CI LB	95% CI UB
γ_{00}	3.35	0.99	244.49	3.39	$p < 0.001$	1.42	5.27
γ_{10}	0.53	0.06	254.14	9.00	$p < 0.001$	0.42	0.65
γ_{20}	0.79	0.21	256.25	3.72	$p < 0.001$	0.37	1.19
γ_{30}	0.19	0.06	256.87	3.39	$p < 0.001$	0.08	0.31
γ_{40}	−0.37	0.17	256.63	−2.19	$p < 0.05$	−0.71	−0.04
γ_{50}	−0.39	0.18	250.04	−2.20	$p < 0.05$	−0.73	−0.05

TABLE X. Overview of all models and their statistical parameters.

	M0	M1	M3	M4	M6a	M6b (final)	Effect size		
Main effects						Raw	Normalized	d_{Cohen}	f^2
Intercept	8.42***	4.88***	1.93*	0.92	3.17**	3.35***			
Pretest	X	0.69***	0.60***	0.56***	0.54***	0.53***	0.45***	1.31	0.39
Theory	X	X	1.12***	0.91***	0.75***	0.79***	0.18***	0.55	0.06
P-Grade	X	X	X	0.20***	0.19**	0.19***	0.17***	0.48	0.06
ICL	X	X	X	X	−0.51**	−0.37*	−0.11*	−0.32	0.02
ECL	X	X	X	X	X	−0.39*	−0.10*	−0.31	0.02
Variance									
Level 1	10.60	7.67	6.87	6.65	6.42	6.35			
Level 2	3.58	1.02	1.03	0.84	0.79	0.72			
Model fit parameters									
BIC	1415.8	1322.1	1299.3	1292.9	1288.2	1288.8			
R^2 (fixed effects)	0	0.35	0.42	0.45	0.48	0.49			
R^2 (total)	0.25	0.42	0.50	0.51	0.53	0.54			

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

VI. DISCUSSION

It seems that both interventions are suitable for improving the conceptual understanding of Newtonian dynamics. It is apparently not relevant whether students theoretically consider the forces and calculate the motion using computational physics or whether, conversely, they experimentally measure the kinematic data and deduce the forces from there. The more important part is that authentic motions with friction are discussed, where cognitive conflicts arise, and students' ideas are put to the test. Students should actively think about and discuss the relationships between the relevant quantities—regardless of the specific software they use.

Various factors that influence the post-test score could be identified: the score in the pretest, the interest in theoretical physics, the cognitive load, and the physics grade. It is not surprising that a higher score in the pretest leads, on average, to a higher score in the post-test, especially since the interventions took place after the lessons in school and lasted around 3.5 h.

Other results are more interesting. For example, students who are more interested in theoretical physics learn more in both interventions, even when the pretest score is controlled. The interventions do not significantly favor the students whose interests are more aligned with the approach, though. It was expected that more theory-interested students learn more in the computational modeling intervention, whereas more experiment-interested students would be more successful in the video motion analysis intervention. This could not be observed.

Knowledge in mathematics does not seem to be necessary to be successful in gaining conceptual understanding of Newtonian dynamics using either computational modeling or video motion analysis. The last grade in mathematics, which is used here, is not a perfect measure for the general math skills. Due to time constraints in the test instrument, it must serve as a sufficiently good proxy. The fact that the math grade does not influence the learning outcome is welcomed since the way the software was used here was supposed to redirect the focus from the necessary mathematics to the relevant physics. The prior physics grade, however, does influence the learning outcome. This can be attributed to the fact that the reported physics grade is from the first semester of the school year where kinematics is traditionally taught in the state of Hesse, where the interventions took place. The interventions took place in the second semester. The ability to interpret graphs is an important skill to be successful using computational modeling and video motion analysis software, which explains the influence of the physics grade.

From a sociocognitive perspective on conceptual change, the perceived relevance of the subject should increase engagement and the construction of new schemata and therefore improve conceptual change. When included in the model without the interest in theoretical physics, it does improve learning gain. The two are highly correlated, though, and the latter explains the observed variance better than the former. Without any other variables included, also the enjoyment of the intervention increases learning gain, while the rating of the used software has no influence on the post-test score.

As expected [46,49,52], the cognitive load (intrinsic and extraneous) negatively influences learning gain. The germane cognitive load did not explain a significant amount of variance in the data, though. Here, a positive influence would be expected from theoretical considerations, which is not observable. The computer affinity did not have an influence on the post-test score, which is positive since the intervention was not supposed to favor students who use a computer more often in their spare time.

The results found here relate to and expand on prior research. Previous findings either concerned a different type of software, university instead of school students, or a different field of physics, though. So far, no effects on the learning gain of equation-based modeling programs have been investigated. In this respect, our results are novel. They are, however, comparable to results in those different areas. Various positive effects, like an increase in correct argumentation [80] and learning gain [97], were previously observed. Here, an increase in learning was shown with a large effect size—and this after the topic had already been taught in school lessons. This observed effect is larger than previously reported effects, where graphical modeling software was used [98]. It was already shown that using computational modeling software is well received by students [36,47]. This was observed here, as well. In addition, the software appears to be easier to learn than was observed with other software [39]. In this study, unlike most other studies, computational modeling was not compared with "conventional" teaching, though, but with another type of software use—video motion analysis, which also proved to be a successful method of learning Newtonian dynamics.

In the area of video analysis, positive results have been observed so far mostly in kinematics (interpretation of graphs [43], vectors [99], conceptual understanding of uniform and accelerated motion, and cognitive load [46,52]). Our results suggest that these positive results appear to be transferable to the area of dynamics.

As always, there are certain caveats to the interpretation of the collected data. Standard procedures were conducted to ensure objectivity, reliability, and validity. As is typical of a comparative study in pre-post design, the challenge is to ensure that the learning gain is only due to the intervention and not to an improvement in test-taking skills or other training between the two tests. Although there is never absolute certainty, there is reason to assume that the learning gain is due to the intervention alone.

First, an improvement due to test training is minimal in the FCI [69]. Since the test used here is similar to the FCI, this should also be the case in this study. Second, a control scale was added to show possible improvements that cannot be explained by an intervention (see Sec. IVA 1). The scale consists of items concerning Newton's third law. No improvement in the conceptual understanding of Newton's third law was expected since this topic was not part of the intervention. One item in the scale did in fact improve significantly, though. The item from the FCI asks for the forces acting on a car and a truck, while the car pushes the truck with a constant velocity. A similar item, where the car and truck both accelerate, did not improve (the score even decreased—but not significantly). The most plausible explanation for the improvement of the first item is the following. Students improved their conceptual understanding of Newton's first law. They learned that the constant velocity of an object is always explained by the sum of all acting forces being zero. They wrongly transferred this knowledge to answer the mentioned item correctly. This is confirmed by the second item, where they do not improve since the car accelerates. If the test-taking skills improved or if they somehow increased their conceptual understanding of Newton's third law, the score in all four items should be increasing. Since only the result of one item has improved and an explanation can be found for this, it is plausible to assume that the overall improvement in Newton's first two laws between pre- and post-test is, in fact, due to the intervention.

Unfortunately, the reduced sample size due to the Covid-19 pandemic reduces the generalizability of the results. Nevertheless, the work provides starting points for further research and is also relevant for physics lessons in schools.

We argue that both investigated methods are beneficial to learn the basic relations in Newtonian dynamics. There are not many differences regarding the results between the two methods. Due to the importance of computational modeling for science in general, and the fact that software such as "Newton-II" seems to be easy enough to use for students without a lengthy introduction, a case for increased use of computational modeling in schools can be made. We showed that video motion analysis on the other hand can be used not only to learn kinematics but also dynamics. It would be beneficial to use the software during the whole year of mechanics lessons regularly since it is known to be beneficial in kinematics (e.g., conceptual understanding of graphs) and the students get used to the software making its use more efficient over time.

A possible next step for a subsequent study would be a field study with the knowledge gained from our analysis and to aim for a higher external validity. Teaching Newtonian dynamics with the use of suitable software to discuss authentic motions with frictional forces that challenge the existing preconceptions is hypothesized to be successful in a school setting as well. Another interesting approach would be to analyze the learning process of students in such a setting with the sociocultural perspective on conceptual change.

ACKNOWLEDGMENTS

We thank the charitable GIERSCH foundation for the financial support of the Physics School Lab of the University of Frankfurt, where the study was conducted.

[1] N. Guicciardini, *Reading the Principia. The debate on Newton's mathematical methods for natural philosophy from 1687 to 1736* (Cambridge University Press, Cambridge, England, 2003).

[2] T. Wilhelm, Verständnis der newtonschen Mechanik bei bayerischen Elftklässlern – Ergebnisse beim Test "Force Concept Inventory" in herkömmlichen Klassen und im Würzburger Kinematik-/Dynamikunterricht, PhyDid A **2**, 47 (2005), http://phydid.physik.fu-berlin.de/index.php/phydid/article/view/32/32.

[3] H. Schecker and T. Wilhelm, Schülervorstellungen in der Mechanik, in *Schülervorstellungen und Physikunterricht*, edited by H. Schecker, T. Wilhelm, M. Hopf, and R. Duit (Springer Berlin Heidelberg, Berlin, Heidelberg, 2018), p. 63, 10.1007/978-3-662-57270-2_4.

[4] P. Potvin, The coexistence claim and its possible implications for success in teaching for conceptual "change", Eur J. Sci. Math. Educ. **5**, 55 (2017).

[5] G. J. Posner *et al.*, Accommodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 211 (1982).

[6] I. A. Halloun and D. Hestenes, Common sense concepts about motion, Am. J. Phys. **53**, 1056 (1985).

[7] D. Hillmayr *et al.*, The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis, Comput. Educ. **153**, 103897 (2020).

[8] Y.-T. Sung, K.-E. Chang, and T.-C. Liu, The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis, Comput. Educ. **94**, 252 (2016).

[9] E. R. Carnes, J. S. Lindbeck, and C. F. Griffin, Effects of group size and advance organizers on learning parameters when using microcomputer tutorials in kinematics, J. Res. Sci. Teach. **24**, 781 (1987).

[10] D. Hillmayr, F. Reinhold, L. Ziernwald, and K. Reiss, *Digitale Medien im mathematisch-naturwissenschaftlichen Unterricht der Sekundarstufe. Einsatzmöglichkeiten, Umsetzung und Wirksamkeit* (Waxmann, Münster, 2017), 10.25656/01:15482.

[11] J. Weber and T. Wilhelm, The benefit of computational modelling in physics teaching: A historical overview, Eur. J. Phys. **41**, 034003 (2020).

[12] H. Schecker, The didactic potential of computer aided modeling for physics education, in *Advanced Educational Technologies for Mathematics and Science*, edited by D. L. Ferguson (Springer Berlin Heidelberg, Berlin, Heidelberg, 1993), p. 165.

[13] S. Mikelskis-Seifert and L. Kasper, Modellieren in der Physik, im Alltag und im Unterricht. Hintergründe und unterrichtliche Orientierung zum Thema Modelle, NiU-Phys. **22**, 4 (2011).

[14] M. Suleder, Eine kurze Geschichte der Videoanalyse, Plus Lucis **28**, 4 (2020), https://www.pluslucis.org/ZeitschriftenArchiv/2020-1_PL.pdf.

[15] P. Potvin *et al.*, Models of conceptual change in science learning: establishing an exhaustive inventory based on support given by articles published in major journals, Stud. Sci. Educ. **56**, 157 (2020).

[16] P. Potvin *et al.*, Coexistence of misconceptions and scientific conceptions in chemistry professors: A mental chronometry and fMRI study, Front. Educ. **5** (2020).

[17] P. Potvin *et al.*, Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference, Int. J. Sci. Math. Educ. **13**, 21 (2015).

[18] A. A. diSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[19] S. Vosniadou, Capturing and modeling the process of conceptual change, Learn. Instr. **4**, 45 (1994).

[20] C. Chinn, A critique of social constructivist explanations of knowledge change, in *Perspectives on Conceptual Change. Multiple Ways to Understand Knowing and Learning in a Complex World*, edited by B. J. Guzzetti (Taylor and Francis, Mahwah, 2013).

[21] R. Driver *et al.*, Constructing scientific knowledge in the classroom, Educ. Res. **23**, 5 (1994).

[22] P. Cobb, Where is the mind? Constructivist and sociocultural perspectives on mathematical development, Educ. Res. **23**, 13 (1994).

[23] M. Gregoire, Is It a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change, Educ. Psychol. Rev. **15**, 147 (2003).

[24] P. R. Pintrich, R. W. Marx, and R. A. Boyle, Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change, Rev. Educ. Res. **63**, 167 (1993).

[25] L. S. Vygostky *et al.*, *Mind in Society* (Harvard University Press, Cambridge, MA, 1980), 10.2307/j.ctvjf9vz4.

[26] L. Mason, Introduction: Bridging the cognitive and sociocultural approaches in research on conceptual change: Is it feasible? Educ. Psychol. **42**, 1 (2007).

[27] A. Sfard, On two metaphors for learning and the dangers of choosing just one, Educ. Res. **27**, 4 (1998).

[28] J. V. Wertsch, *Mind as Action* (Oxford University Press, New York, 1998), 10.1093/acprof:oso/9780195117530.001.0001.

[29] J. Lave, *Cognition in Practice Mind, Mathematics and Culture in Everyday* (Cambridge University Press, Cambridge, England, 1988), 10.1017/CBO9780511609268.

[30] H. Schecker, *Physik—Modellieren. Grafikorientierte Modellbildungssysteme im Physikunterricht* (Klett, Stuttgart, 1998).

[31] L. Hucke, Handlungsregulation und Wissenserwerb in traditionellen und computergestützten Experimenten des physikalischen Praktikums, Dissertation, Logos-Verlag, Berlin, 2000, Vol. **8**, https://d-nb.info/960396934/34.

[32] F. Sander, H. Schecker, and H. Niedderer, Wirkungen des Einsatzes grafikorientierter Modellbildung im physikalischen Praktikum, Z. Didakt. Naturwiss. **7**, 147 (2002).

[33] H. Schecker *et al.*, Abschlussbericht zum DFG-Projekt—Physiklernen mit Modellbildungssystemen. Förderung physikalischer Kompetenz und systemischen Denkens durch computergestützte Modellbildungssysteme, 1999.

[34] J. Benacka, Projectile general motion in a vacuum and a spreadsheet simulation, Phys. Educ. **50**, 58 (2015).

[35] J. Benacka, Spreadsheet application showing the proper elevation angle, points of shot and impact of a projectile, Phys. Educ. **50**, 342 (2015).

[36] J. Benacka, Numerical modelling with spreadsheets as a means to promote STEM to high school students, Eurasia J. Math. Sci. Technol. Educ. **12** (2016).

[37] D. Scherer, P. Dubois, and B. Sherwood, VPython: 3D interactive scientific graphics for students, Comput. Sci. Eng. **2**, 56 (2000).

[38] M. D. Caballero, M. A. Kohlmyer, and M. F. Schatz, Implementing and assessing computational modeling in introductory mechanics, Phys. Rev. ST Phys. Educ. Res. **8** (2012).

[39] M. D. Caballero *et al.*, Integrating numerical computation into the modeling instruction curriculum, Phys. Teach. **52**, 38 (2014).

[40] P. W. Laws and H. Pfister, Using digital video analysis in introductory mechanics projects, Phys. Teach. **36**, 282 (1998).

[41] T. Wilhelm and T. Trefzger, Erhebung zum Computereinsatz bei Physik-Gymnasiallehrern, PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung (2010), https://ojs.dpg-physik.de/index.php/phydid-b/article/view/109/119.

[42] M. Wenzel, Computereinsatz in Schule und Schülerlabor, Dissertation, Logos-Verlag, Berlin, 2018, Vol. **251**, https://www.logos-verlag.de/ebooks/OA/978-3-8325-4659-5.pdf.

[43] R. J. Beichner, The impact of video motion analysis on kinematics graph interpretation skills, Am. J. Phys. **64**, 1272 (1996).

[44] P. Hockicko, B. Trpišová, and J. Ondruš, Correcting Students' Misconceptions about Automobile Braking Distances and Video Analysis Using Interactive Program Tracker, J. Sci. Educ. Technol. **23**, 763 (2014).

[45] L. Escalada, R. Grabhorn, and D. Zollman, Applications of interactive digital video in a physics classroom, J. Educ. Multimed. Hypermedia **5**, 73 (1996).

[46] S. Becker *et al.*, Förderung von Konzeptverständnis und Repräsentationskompetenz durch Tablet-PC-gestützte Videoanalyse, Z. Didakt. Naturwiss. **25**, 1 (2019).

[47] T. Wilhelm, Konzeption und Evaluation eines Kinematik/Dynamik-Lehrgangs zur Veränderung von Schülervorstellungen mit Hilfe dynamisch ikonischer Repräsentationen und graphischer Modellbildung, Dissertation, Logos-Verlag, Berlin, 2005, Vol. **46**, https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/3310.

[48] K. Hochberg *et al.*, Using smartphones as experimental tools—a follow-up: Cognitive effects by video analysis and reduction of cognitive load by multiple representations, J. Sci. Educ. Technol. **29**, 303 (2020).

[49] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988).

[50] J. J. G. van Merriënboer and J. Sweller, Cognitive load theory and complex learning: Recent developments and future directions, Educ. Psychol. Rev. **17**, 147 (2005).

[51] J. Leppink and A. van den Heuvel, The evolution of cognitive load theory and its application to medical education, Perspect. Med. Educ. **4**, 119 (2015).

[52] S. Becker *et al.*, Investigating dynamic visualizations of multiple representations using mobile video analysis in physics lessons, Z. Didakt. Naturwiss. **26**, 123 (2020).

[53] J. Weber, Mathematische Modellbildung und Videoanalyse zum Lernen der Newtonschen Dynamik im Vergleich, Dissertation, Logos-Verlag, Berlin, 2022, Vol. **345**, 10.30819/5566.

[54] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber, *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (Springer Gabler, Berlin, Heidelberg, 2016), 10.1007/978-3-662-46076-4.

[55] J. M. Graham, Congeneric and (essentially) tau-equivalent estimates of score reliability, Educ. Psychol. Meas. **66**, 930 (2006).

[56] E. Cho, Making reliability reliable, Organ. Res. Meth. **19**, 651 (2016).

[57] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory* (McGraw-Hill, New York, NY, 2008).

[58] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[59] A. Savinainen and P. Scott, The Force Concept Inventory: A tool for monitoring student learning, Phys. Educ. **37**, 45 (2002).

[60] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[61] D. Huffman and P. Heller, What does the force concept inventory actually measure? Phys. Teach. **33**, 138 (1995).

[62] J. Gerdes and H. Schecker, Der Force Concept Inventory. Ein diagnostischer Test zu Schülervorstellungen in der Mechanik, MNU **52**, 283 (1999).

[63] A. M. Just, C. von Aufschnaiter, and A. Vorholzer, Effects of conceptual and contextual task characteristics on students' activation of mechanics conceptions, Eur. J. Phys. **42**, 025702 (2021).

[64] T. Wilhelm, Vektorverständnis von Studienanfängern, in *Didaktik der Physik. Regensburg 2007*, edited by V. Nordmeier, A. Oberländer, and H. Grötzebauch (Lehmanns Media LOB.de, Berlin, 2007).

[65] S. Flores, S. E. Kanim, and C. H. Kautz, Student use of vectors in introductory mechanics, Am. J. Phys. **72**, 460 (2004).

[66] J. Warren, *Understanding Force* (John Murray, London, 1979).

[67] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[68] D. Heuer and T. Wilhelm, Aristoteles siegt immer noch über Newton, MNU **50**, 280 (1997).

[69] H. Schecker and J. Gerdes, Messung der Konzeptualisierungsfähigkeit in der Mechanik - Zur Aussagekraft des Force Concept Inventory, Z. Didakt. Naturwiss. **5**, 75 (1999).

[70] M. Laukenmann *et al.*, Eine Untersuchung zum Einfluss emotionaler Faktoren auf das Lernen im Physikunterricht, Z. Didakt. Naturwiss. **6**, 139 (2000).

[71] J. Leppink *et al.*, Development of an instrument for measuring different types of cognitive load, Behav. Res. Methods **45**, 1058 (2013).

[72] S. Becker-Genshow, Lernwirksamkeit einer Tablet-PC-gestützten Videoanalyse im Mechanikunterricht der Sekundarstufe - Ergebnisse quasi-experimenteller Feldstudien zu Konzeptverständnis, kognitiver Belastung und induzierten Emotionen, Dissertation, Technische Universität Kaiserslautern, 2020.

[73] J. Leppink *et al.*, Effects of pairs of problems and examples on task performance and different types of cognitive load, Learn. Instr. **30**, 32 (2014).

[74] M. Thees *et al.*, Comparing two subjective rating scales assessing cognitive load during technology-enhanced STEM laboratory courses, Front. Educ. **6** (2021).

[75] R. Eisinga, M. t. Grotenhuis, and B. Pelzer, The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int. J. Public Health **58**, 637 (2013).

[76] A. Helmke, Determinanten der Schulleistung: Forschungsstand und Forschungsdefizite, in *Empirische Pädagogik 1970-1990. Eine Bestandsaufnahme der Forschung in der Bundesrepublik Deutschland.*, edited by K. Ingenkamp, R. Jäger, H. Petillon, and B. Wolf (Deutscher Studien Verlag, Weinheim, 1992), p. 595.

[77] M. Kunter *et al.*, *PISA 2000: Dokumentation der Erhebungsinstrumente* (Max-Planck-Institut für Bildungsforschung, Berlin, 2002), Vol. 72.

[78] C. Brell, Lernmedien und Lernerfolg—reale und virtuelle Materialien im Physikunterricht. Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulation und IBE, Dissertation Logos-Verlag, Berlin, 2008, Vol. **74**.

[79] J. Chen *et al.*, The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: A meta-analysis, Rev. Educ. Res. **88**, 799 (2018).

[80] H. Schecker and J. Gerdes, Interviews über Experimente zu Bewegungsvorgängen, Z. Didakt. Naturwiss. **4**, 61 (1998).

[81] S. Becker, P. Klein, and J. Kuhn, Video analysis on tablet computers to investigate effects of air resistance, Phys. Teach. **54**, 440 (2016).

[82] J. Leisen and M. Neffgen, Modellbildungspraktikum: Fall von Körpern in Luft, PdN-PhiS **48**, 7 (1999).

[83] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (L. Erlbaum Associates, Hillsdale, NJ, 1988), 10.4324/9780203771587.

[84] J. Kuhn, Mehrebenenanalyse am Beispiel der Lernwirkung von Aufgaben, in *Methoden in der naturwissenschaftsdidaktischen Forschung*, edited by D. Krüger, I. Parchmann, and H. Schecker (Springer Spektrum, Berlin, Heidelberg, 2014), p. 297, 10.1007/978-3-642-37827-0_24.

[85] S. W. Raudenbush and A. S. Bryk, *Hierarchical Linear Models. Applications and Data Analysis Methods* (SAGE Publ., Thousand Oaks, CA, 2010), Vol. 1.

[86] H. Ditton, *Mehrebenenanalyse. Grundlagen und Anwendungen des hierarchisch linearen Modells* (Juventa-Verl., Weinheim, 1998).

[87] J. J. Hox, *Multilevel Analysis. Techniques and Applications* (Routledge, New York, NY, 2010).

[88] M. Gollwitzer *et al.*, Effectiveness of the Vienna social competence training on explicit and implicit aggression, Eur. J. Psychol. Assess. **23**, 150 (2007).

[89] C. J. M. Maas and J. J. Hox, The influence of violations of assumptions on multilevel parameter estimates and their standard errors, Comput. Stat. Data Anal. **46**, 427 (2004).

[90] C. J. M. Maas and J. J. Hox, Sufficient sample sizes for multilevel modeling, Methodology **1**, 86 (2005).

[91] J.-P. Burde, Konzeption und Evaluation eines Unterrichtskonzepts zu einfachen Stromkreisen auf Basis des Elektronengasmodells, Dissertation Logos-Verlag, Berlin, 2018, Vol. **259**, 10.30819/4726.

[92] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (University of Chicago Press, Chicago, 1980).

[93] H. Wu and S.-O. Leung, Can Likert scales be treated as interval scales?—A simulation study, J. Soc. Serv. Res. **43**, 527 (2017).

[94] W. J. Browne, Applying MCMC methods to multi-level models, Dissertation, University of Bath, 1998.

[95] A. Field, *Discovering Statistics Using SPSS (and Sex, Drugs and Rock'n'Roll)* (SAGE Publication, London, 2007).

[96] J. Hartig and K. Rakoczy, Mehrebenenanalyse - multilevel analysis, in *Handbuch Statistik, Methoden und Evaluation*, edited by H. Holling (Hogrefe, Göttingen, 2010), p. 538.

[97] P. W. Irving, M. J. Obsniuk, and M. D. Caballero, P3: a practice focused learning environment, Eur. J. Phys. **38**, 055701 (2017).

[98] F. Sander, *Verbindung von Theorie und Experiment im physikalischen Praktikum. Eine empirische Untersuchung zum handlungsbezogenen Vorverständnis und dem Einsatz grafikorientierter Modellbildung im Praktikum*, Zugl.: Bremen, Univ., Diss., 1999 (Logos-Verl., Berlin, 2000), Vol. 13.

[99] P. Klein, J. Kuhn, and A. Müller, Förderung von Repräsentationskompetenz und Experimentbezug in den vorlesungsbegleitenden Übungen zur Experimentalphysik, Z. Didakt. Naturwiss. **24**, 17 (2018).