Assessing the effectiveness of hands-on lessons, online lessons or combinations of both for tackling students' misconceptions about Newton's first law

Maaike Otten

University of Twente

Supervised by:

S.A.N. van Riesen, Msc.

Dr. A.H. Gijlers

Enschede, 2016

Acknowledgment

I would like to thank a lot of people who have supported and helped me throughout my final project for the master Psychology, Learning sciences. Firstly, I would like to give thanks to my supervisors Siswa van Riesen and Hannie Gijlers. I am glad that I met Siswa van Riesen and visit her to talk about opportunities for research projects at the IST department. I liked the open attitude and thinking along about possible topics for the research projects, which also characterised the support during the project by offering freedom and also pointing attention to possible areas for improvement. I am glad that she introduced me to Hannie Gijlers, who could quickly answer my questions and provided valuable feedback to bring my thesis to the next level.

Secondly, I am grateful that I was given the opportunity to be part of the project Miss to Hit. I liked that I could contribute to the project and enjoyed the first project meetings. The project Miss to Hit also enabled me to pre-test the instruments and materials at NEMO Science Museum, which was a nice location for the pre-testing.

Thirdly, I would also like to thank Veluwscollege Mheenpark and Veluwscollege Walterbosch for conducting my research at their school and especially the teachers and students that participated in this research.

Moreover, I am grateful that I could use the inquiry spaces from Go-lab and the racing tracks for the lessons. Special thanks to the people of the Go-lab platform for answering questions and to make it possible to integrate labs in the inquiry spaces.

Finally, the process of meeting one person who can help me to get in touch with someone else, seemed to portray my journey of the whole thesis. When I talked about my thesis to friends or fellow students, they could offer me great tips for my thesis and also suggestions for meeting other people who could help me. I am grateful for the advice of family, friends and fellow students but also for the feedback of the several kind strangers.

Thanks to all of you.

Maaike Otten, August 2016

Abstract

Students hold on misconceptions about Newton's first law which are tenacious and are often hard to change. Scientists are debating what the most effective method is to tackle those misconceptions. This research inquires what the most effective method is, comparing hands-on lessons, online lessons or combinations of both, to tackle students' misconceptions about Newton's first law. Secondary school students at the third grade level (*N*= 127) were assigned to four different conditions (1. hands-on lessons, 2. online lessons, 3. online lesson and then hands-on lesson, 4. hand-on lesson and then online lesson). Science curiosity scales and four-tier tests measuring (mis)conceptions and confidence were used as instruments. This study showed that no method or combination of methods was superior to another for student's performances and confidence about Newton's firs law and altering students' science curiosity. The findings contribute to the scientific debate about which method is the most effective for acquiring conceptual change and stimulating science curiosity. The study offers educational practises and initiatives for promoting science curiosity insights if preferred instruction methods would exist. Furthermore, by implementing a four-tier test including open questions which gains insights in students' confidence, it contributes as a methodological development of the four-tier diagnostic multiple choice test for research practices in the domain of misconceptions.

Keywords: hands-on lesson, online lesson, misconceptions, confidence, science curiosity

Introduction

Technologies are gaining an increasingly important role in modern society, which as a result demand more technical labour instead of unskilled workers (Wysocki, McDonald, Fanto & McEwen, 2013). In the growth of modern society, science, technology, engineering and mathematics (STEM) jobs are necessary for growth (Wysocki et al., 2013). Students' viewpoints on science impact their future career choices (Jocz, Zhai & Tan, 2014; Morais, 2015). Besides that, Jocz et al. (2014) suggest that disinterest in science may impede students to become scientific literate citizens who can handle socioscientific issues. Unfortunately, children tend to show negative perceptions towards science (Morais, 2015), this is especially the case for girls (Heard, Divall, & Johnson, 2000; Smiley, 2011). Children's opinions about science are rooted early in their education and are difficult to change later on (Kermani & Aldemir, 2015; Morais, 2015).

On a positive note, early multidisciplinary science instruction raises awareness and interests towards science and promotes their overall school performance even years later (Kermani & Aldemir, 2015). Besides that, young children are intrinsically interested in science because they are eager to learn about the world around them and the underlying causes, processes and mechanisms of phenomena (Zacharia, Loizou & Papaevripidou, 2012).

However, children's interests or own investigations can cause problems for their conceptions about scientific phenomena. Children attempt to make sense of the world around them and set up their own workable theories to clarify their experiences (Garbett, 2003). Misconceptions are more the rule than the exception (Pompea, Dokter, Walker, & Sparks, 2007). Misconceptions are tenacious and are often hard to change (Boyes, 1988; Garbett, 2003; Digisi & Yore, 1992; Unal, 2008). Therefore, teaching in science will not be offering explanations about entirely new topics but would rather be altering prior intuitions and trying to change or build upon those misconceptions (Levitt, 2001; Mayer, 2008). Doubting about own misconceptions is the first step of conceptual change (Posner , Strike, Hewson & Gertzog, 1982). Confidence measures can help to determine those shifts in conceptual change (Gurel, Eryılmaz & McDermott, 2015). Students often have misconceptions about motion which underlines a lack of Newton's scientific conception about motion (e.g. Digisi & Yore 1992;

Kozhevnikov, Gurlitt & Kozhevnikov, 2013; Mayer, 2008; Vosniadou, Ioannides, Dimitrakopoulou, Papademetriou, 2001). Teachers and textbooks also hold misconceptions, which in turn can even accumulate students' misconceptions (Bursal, 2012; King, 2010; Bulunuz & Jarret, 2010; Garbett, 2003; Treagust & Duit, 2008). Being aware of misconceptions can contribute to the adaption of instructions to help students acquire scientific conceptions (Pine, Messer & St. John, 2001).

Additional, research shows that traditional methods, involving primarily lectures alone, are not able to tackle students' misconceptions effectively but that hands-on lessons, in which would be experimented with physical material, are considered as more effective and enjoyable in order to clarify misconceptions (Marinopoulos & Stavridou, 2002; Hadzigeorgiou, 2001; Weaver, 1998). Furthermore, hands-on lessons promote students' interest in science (Klahr, Triona & Williams, 2007; Unal, 2008; Zardetto-Smith, Mu, Phelps, Houtz & Royeen, 2002). Though hands-on lessons also hold disadvantages such as students having problems to recognise underlying principles and phenomena (Lazonder & Ehrenhard, 2014; Smith & Puntambekar, 2010). Online lessons can provide solutions, such as showing phenomena that are not readily visible or manipulated (Klahr et al., 2007; De Jong, Linn & Zacharia, 2013; Wysocki et al., 2013). Nevertheless, online lessons also hold drawbacks such as missing the tactile cues that can be derived from physical experimentation (Klahr et al., 2007; Lazonder & Ehrenhard, 2014; Zacharia et al., 2012). Combinations of hands-on lessons and online lessons are proposed in order to use the best of both worlds (e.g. De Jong et al., 2013; Olympiou & Zacharia, 2011). The literature do not show consensus about what type of method (comparing hands-on lessons, online lessons or combinations of both) the most affordances has (Klahr et al., 2007; Lazonder & Ehrenhard, 2014; Zacharia, 2007). Whereas science education need to know when, what and how to teach, using hands-on and or online material in order to optimize students achievements, because there is a lack of a framework in science curricula about acquiring scientific conceptions using hands-on and online material (Zacharia et al., 2008; Zacharia et al., 2012).

The current study inquires what the most effective method is, comparing hands-on lessons, online lessons or combinations of both, to tackle students' misconceptions about Newton's first law. (Klahr et al., 2007; Lazonder & Ehrenhard, 2014; Zacharia, 2007). Research into different domains, with different types of students and instructional goals will add to the body of knowledge about the effectiveness of hands-on lessons, online lessons and combinations of both (Klahr et al., 2007). This study also measures students’ long-term recall, confidence in their (mis)conception, and students’ interest in science. Misconceptions are very resistant to change thus it will be worthwhile to detect even small shifts in conceptual change. Confidence measures can help to determine shifts in conceptual change in nature and strength (Gurel et al., 2015). Measuring students’ (mis)conceptions and confidence at different times would also enable to measure the degree of tenacity of the (mis)conceptions. Taking into account affective factors, such as science curiosity and confidence, besides the performances measures when conducting research about conceptual change is recommended by Duit and Treagust (2003). Furthermore, previous studies primarily measure conceptual understanding but do not pay attention to science curiosity (De Jong et al., 2013). Even though science curiosity is an interesting topic in the current society to examine because of the need for more STEM jobs and the influence of students’ interest on their future career choices (Jocz et al., 2014; Morais, 2015; Wysocki et al., 2013).

Theoretical framework

Misconceptions

Levin and Druyan (1993, p.1572) describe misconceptions as “intuitive conceptions incompatible with scientific knowledge and [...] particularly resistant to change through schooling and other daily experiences”. Schmidt, Spaaij and De Grave (1988) add to these characteristics of misconceptions that those views emerge kind of spontaneously in an early stage of development through prior experiences. Synonyms that are often used for the term misconceptions are; naive views, alternative conceptions, children’s ideas, preconceptions or common sense conceptions (Clement, 1982; Garbett, 2003; Gurel et al., 2015; Unal, 2008).

Students hold misconceptions about a lot of topics in science. Mayer (2008) summarize that the known misconceptions are about gravity, acceleration, motion, density, living versus non-living, chemical equilibrium, energy, the sky and the earth as a cosmic body (Mayer, 2008). Other misconceptions that are mentioned are for example; balance (Zacharia et al., 2012; Pine et al., 2001) and floating concepts and rules (Unal, 2008). Many researchers refer to misconceptions about motion as an example of possible misconceptions (e.g. Digisi & Yore 1992; Kozhevnikov et al., 2013; Vosniadou et al., 2001). Related to the misconceptions of motion are misconceptions about vision and light (Winer, Rader & Cottrell, 2003; Pompea et al, 2007). Specifically, misconceptions about motion for example include that static objects cannot exert forces, an unbalanced force is necessary to keep an object moving with constant velocity and that objects shoot through a curved tube will follow a circular path instead of going in a straight line (Clement, 1982; McCloskey, 1983; Muller, Bewes, Sharma & Reimann, 2008). Abstract concepts, such as forces, are considered as the hardest topics for students to understand (Bulunuz, Jarret & Bulunuz, 2009; Pine et al., 2001; Unal, 2008). This is especially the case for misconceptions about motion because students must discriminate between concrete, observable events and unobservable properties that cause them (Unal, 2008). These misconceptions about motion are tenacious (McCloskey, 1983; Mildenhall, & Williams, 2001; Treagust & Duit, 2008). For example, Hynd and Alverman (1986), as cited in Digisi and Yore (1992), show that even when misconceptions are explicitly stated, students will remember the information but still have problems to apply the theory to new situations. The theory of impetus underlines the misconceptions of motion (Mayer, 2008). This theory involves that a moving object acquires some force or impetus that keeps it moving until the impetus gradually will dissipate. Accordingly, students who believe the theory of impetus supports the idea that motion requires a force. Despite the fact that Newton's first law is considered as the scientific conceptions about motion (Mayer, 2008). Newton's first law which encompasses that an object in motion will continue in his current state until some external force acts upon it. Besides that, an object in rest will also stay in rest until some external force acts upon it. Thus, objects do not require any force to continue moving at a constant speed but an external force is necessary to change the velocity and direction of a moving object (Mayer, 2008). The current study aims to tackle those misconceptions about motion which are underlined by a lack of knowledge of Newton's first law because those misconceptions frequently occur and are tenacious. Moreover, Newton's scientific conceptions are the base of classical physics and usually are the groundwork of introductory physics curricula (Galili & Tseitlin, 2003).

Several authors propose approaches to deal with misconceptions (Mildenhall & Williams, 2001; Pompea et al., 2007). In general, there is a consensus that misconceptions should be identified and student's awareness should be raised for their own misconceptions (Mildenhall & Williams, 2001; Pompea ea al., 2007; Vosniadou et al., 2001). Those misconceptions can be challenged by introducing anomalous data (Bulunuz, Jarret & Bulunuz, 2009; Chinn & Brewer, 1993; Mayer, 2008), while a few also suggest that the preconceptions can be used as foundations on which to build other concepts by the use of analogy or bridging (Clement, 1982; Mildenhall & Williams, 2001). The former is more in line with the widely adhered conceptual-change theories than the latter (Mildenhall & Williams, 2001). Conceptual change theory have its roots in Piaget's ideas that people attempt to equilibrate cognitive conflict through the processes of assimilation and accommodation (Bulunuz & Jarret, 2010; Duit & Treagust, 2003). Contradicting experiences can result in a state of disequilibrium in an individual. This can be solved by assimilation; new information will be connected to existing knowledge, or accommodation; existing knowledge will be replaced by a new conception (Bulunuz, Jarret & Bulunuz, 2009; Mayer, 2008). These new conceptions should be intelligible, fruitful and plausible to be able to replace the former misconceptions (Posner et al., 1982). Treagust and Duit (2008) recommend multi-perspectives of conceptual change that address epistemological, ontological and affective domains in order to entirely capture the learning processes. Popular methods for dealing with misconceptions are teaching students appropriate scientific skills, such as control-variable strategy (Acar, 2014, Mayer, 2008) or using an educational strategy in which students adopt a scientific approach, like the predict-observe-explain method and inquiry-based learning (Chinn & Brewer, 1993; De Jong, 2006; Pedaste et al., 2015). Inquiry-based learning was considered as one of the most effective methods for acquiring conceptual change (Kollöffel & De Jong, 2013; Roehrig & Garrow, 2007; So, Cheng, Kong, & Ching, 2014). Moreover, inquiry-based learning can promote science curiosity (So et al., 2014). During inquiry-based learning, students are taking responsibility for their learning and are being actively engaged by doing experiments which are guided by the phases orientation, conceptualization, investigation, conclusion and discussion (De Jong, 2006; Pedaste et al., 2015). Especially, a phase of discussion during lessons is mentioned by many researchers as an important strategy for dealing with misconceptions (Clement, 1982; Vosniadou et al., 2001; Waldrip, Prain & Sellinger, 2013). Furthermore, several kinds of mediums are used to cope with misconceptions, varying from textbooks, concept maps, field trips to computer simulations (Bulunuz & Jarret, 2010; Mildenhall & Williams, 2001).

Hands-on lessons

While considering the diversity of approaches and mediums, hands-on lessons is considered as one of the most popular and effective type of lessons to change misconceptions over the last decades (Bulunuz & Jarret, 2010; Bulunuz, Jarret & Bulunuz, 2009; Hadzigeorgiou, 2001; Wysocki et al, 2013; Weaver, 1998). Hands-on lessons are characterised as an instructional strategy that encourage students to be actively engaged in scientific processes by manipulating objects while using their senses (Inan & Inan, 2015; Klahr et al., 2007).

Hands-on materials have benefits above the use of other mediums for dealing with misconceptions. According to Lazonder and Ehrenhard (2014) physical experimentation have an advantage over the other instructional strategies such as online experimentation when tactile cues that are derived from physical experimentation are not possible to obtain by other senses (Lazonder & Ehrenhard, 2014). The tactile cues such as touching, smelling and manipulating in three-dimensional space are not possible to acquire from an online experience (Lazonder & Ehrenhard, 2014). Zacharia et al. (2012) conclude that hands-on lessons is more beneficial than online lessons when students have incorrect prior knowledge. The modality principle is mentioned as an explanation (Lazonder & Ehrenhard, 2014; Zacharia et al., 2012). This principle involves that when information is processed by multiple modalities, such as visual and tactile cues, it enhances learning more than learning by only one modality. Physical experimentation stimulates learning because it can provide additional sources of brain activation that facilitates retention and retrieval (Klahr et al., 2007; Lazonder & Ehrenhard, 2014). Hands-on lessons are possible in the case of the misconceptions about motion, because of many kinaesthetic opportunities this topic can bring such as pushing and accelerating objects (Carrejo & Reinhartz, 2014). Furthermore, students can gain insights and skills in the world of research when they are doing hands-on lessons (Musacchio, Piangiamore, D'Addezio, Solarino & Eva, 2015; Unal, 2008). Experimenting with physical material requires of students to cope with challenges many scientists face when planning investigations; observations of real-world data over long time spans and troubleshooting (De Jong et al., 2013; Smith & Puntambekar, 2010). Furthermore, hands-on lessons can stimulate interest in science (Klahr et al., 2007; Unal, 2008; Zardetto-Smith et al., 2002). Reasons why hands-on activities are motivating is because students can directly experience and acquire a sense of ownership (Zardetto-Smith et al., 2002). This is especially the case when the activities are relevant to daily life or experience (Weaver, 1998). Using simple materials to challenge students' misconceptions are in line with Piaget's ideas of equilibration (Bulunuz, Jarret & Bulunuz, 2009). Familiarity which can support assimilation in combination with incongruent experiences can encourage disequilibrium.

Online lessons

Hands-on lessons also have drawbacks (Klahr et al., 2007; Lazonder & Ehrenhard, 2014; Wysocki et al., 2013). Therefore, the preference is not always given to hands-on lessons (Klahr et al., 2007; Wysocki et al., 2013). For example, when phenomena are not readily visible or manipulated, such as atoms (Wysocki et al., 2013; Zacharia et al., 2012). Another drawback is that students can fail to recognise and apply the abstract principle underlying a concrete problem or situation while using concrete materials (Gick & Holyoak, 1980; Smith & Puntambekar, 2010).

Online lessons can be a solution for the downsides of hands-on lessons, especially for inquiry-based classrooms (e.g. Smith & Puntambekar, 2010; Zacharia & Olympiou, 2011). Online experimentation can be characterised as virtually interacting or experimenting with dynamic visualisations of the natural or material world (Zacharia & Olympiou, 2011). Those dynamic visualisations of phenomena may show contradictory experiences with student's prior conceptions which as a result can serve as discrepant events that induce students to make conceptual changes (Zacharia & Anderson, 2003). Online lessons can visualise and simplify phenomena that are normally beyond perception or possibility to manipulate (De Jong et al., 2013; Olympia & Zacharia, 2012; Yang, Streveler, Miller, Slotta, Matusovich & Magana, 2012). Furthermore, students may be easier able to recognise principles because online lessons can adapt reality (De Jong et al., 2013). Students' attention can be more directly focused on the targeted phenomena by highlighting salient information or excluding measurement errors and showing a more simplified real-world model (Olympia &

Zacharia, 2012; Smith & Puntambekar, 2010). Besides that, characteristics like the time scale can be adapted in order to make interpretations easier and students can be encouraged to draw conclusions about several observations and linking them to symbolic equations (De Jong et al., 2013). Online lessons have several different affordances than hands-on lessons, such as safety, cost-efficiency, scaffolding, making or repeating measurements quickly and minimisation of errors (De Jong et al., 2013; Smith & Puntambekar, 2010; Triona & Klahr, 2003; Zacharia, 2007; Zacharia et al., 2015). Online lessons show to have a positive impact on students' skills, attitudes and conceptual understanding (Zacharia, 2007). Yang et al. (2012) indicate that participants like the ability to self-pace and flexibility of the schedule, such as repeating material whenever he or she prefers, about online learning. Furthermore, online lessons can provide multiple representations (verbal, numerical, pictorial, conceptual and graphical) and immediate feedback (Olympia & Zacharia, 2012; Smith & Puntambekar, 2010; Zacharia, 2007). Moreover, online lessons are less time consuming because they usually require less setup time and provide results of lengthy investigations instantaneously (De Jong et al., 2013; Smith & Puntambekar, 2010). As a result students are able to perform more experiments within the same amount of time and thereby gather more information or examples (De Jong et al., 2013; Olympia & Zacharia, 2012; Zacharia, 2007).

On the other hand, online lessons also include weaknesses for dealing with misconceptions, like the lack of tactile cues (Lazonder & Ehrenhard, 2014; Zacharia et al., 2012). Moreover, the perceived credibility of the online lessons can impede conceptual change (Lazonder & Ehrenhard, 2014). Students often regard the conflicting data as being unworthy of consideration in order to protect their incorrect initial beliefs (Chinn & Brewer, 1993). For example, students reject the findings by pointing out towards methodological issues (e.g. moving objects in a simulation are not affected by friction) or reinterpret the data in line with their misconceptions (e.g. the findings are correct for an ideal, online world but in real life things work differently)(Lazonder & Ehrenhard, 2014).

Hands-on lessons and online lessons

Scientists do not agree whether hands-on or online lessons are more effective for dealing with students' misconceptions. Several studies indicate that online lessons are equal or more effective than hands-on lessons and situations in which hands-on lessons are most appropriate (De Jong et al., 2013; Klahr et al., 2007; Olympia & Zacharia, 2012; Triona & Klahr, 2003; Zacharia & Olympiou, 2011). For example, Chini, Madsen, Gire, Rebello and Puntambekar's (2012) study indicate that the concept of the lessons and the timing of the post-test influence the results which type of lesson is more beneficial. In their case, the topics force and mechanical advantage do not show any difference between the hands-on or online condition, but the topic 'work' shows better results for the online condition directly after the activity but similarly on the test one week later.

Hands-on lessons and online lessons have both their own pros and cons (Olympia & Zacharia, 2012). Combinations of hands-on lesson and online lessons can build on each other's affordances (De Jong et al., 2013; Zacharia, Olympiou & Papaevripidou, 2008). Any combinations of hands-on lessons and online lessons produce higher learning gains than either of its constituents (Chini et al., 2012; Zacharia, 2007; De Jong et al., 2013; Zacharia et al., 2012; Zacharia et al., 2008). Lazonder and Ehrenhard (2014) suggest that the combination of hands-on and online lessons will produce the best learning outcomes because of the benefits of multiple representations rather than the single representation of only the hands-on lesson or online lesson.

Contradictory findings of studies are found about the order of experimenting with hands-on material or online material. In the studies of Gire et al. (2010) and Smith & Puntambekar (2010) the students who first experiment with real hands-on material and then with the online material score higher on conceptual understanding than students who conduct the research in reverse order. Whereas, other researchers do not find any significant differences between the order of lessons (Chini et al., 2012; Toth, Morrow & Ludvico, 2009). Though Toth et al.'s (2009) findings suggest a little preference for the sequence of online lessons first and then the hands-on lessons. Toth et al. (2009) found a small but not significant advantage for starting with the online experimentation and then subsequently experimenting with hands-on material. Their qualitative data shows that students prefer the condition that starts with online lesson followed by the hands-on lesson.

Research question

The current study aims to answer the following primary research question:
What is the most effective method, comparing hands-on lessons, online lessons or combinations of both (online lesson and then hands-on lesson; hand-on lesson and then online lesson) to tackle students' misconceptions about Newton's first law?

Other secondary questions are taken into account:

1. Which method will decrease students' misconceptions about Newton's first law to scientific conceptions the most, hands-on lessons, online lessons or combinations of both?
2. Which method has the best retention effect in maintaining scientific conceptions of students about Newton's first law over time, hands-on lessons, online lessons or combinations of both?
3. Which method will change students' confidence in their (mis)conceptions about Newton's first law the most, hands-on lessons, online lessons or combinations of both?
4. Which method will increase students' interest in science the most, hands-on lessons, online lessons or combinations of both?

Method

Respondents

The research was conducted on two locations of a school at the HAVO (Higher General Secondary Education) and VWO (Pre-university education) level in The Netherlands. In total participated 183 students, but a number of students were removed from the dataset for the analyses based on the following criteria: Students did not attend or complete all the lessons or assessments of the (mis)conception test; students had inadvertently acquainted with another condition; students who had the online lesson when the online learning environment did not work properly and students who were not willing to fill in their own name on the questionnaires and misconception test. After removing those participants from the dataset, 127 secondary school students in total remained for the analysis (N_{boys} =63 and N_{girls}= 64). Participants who did not fill in the science curiosity questionnaire remained in the sample but their data was assigned as missing for relevant analyses. With the pre-test 19 questionnaires were missing and 5 questionnaires were missing at the post-test 2.

Based on the pre-test of the whole sample (N=183), students were matched to the four different conditions based on their performance on the pre-test. Students were divided in groups by the performances at the pre-test, namely 25% low performing students, 50% average performing students, 25% high performing students. Subsequently, students were randomly assigned to each condition which resulted in the following distribution of participants to the conditions for the analysis (see Table 1): 32 students for the hands- on lesson condition; 34 students for the online lessons condition; 31 students for the online lesson and then hands-on lesson condition and 30 students for the hands-on lesson and then online lesson condition. From the selected sample for the analysis, 63 boys and 64 girls participated. Furthermore, 21 students were at the 3 Havo level and 106 students at the 3 Vwo level.

Table 1
Characteristics participants

	Condition				
	Hands-on	Online	Online & hands-on	Hands-on & online	Total
Performance pre-test *M* (*SD*)	3,39 (1,85)	3,41 (1,50)	3,29 (1,49)	3,40 (1,50)	3,37 (1,57)
Low performer *n*	10	8	8	8	34
Average performer *n*	14	19	17	14	64
High performer *n*	8	7	6	8	29
Gender *n*					
Boy	16	16	17	14	63
Girl	16	18	14	16	64
Level *n*					
3 HAVO	3	6	6	6	21
3 VWO	29	28	25	24	106

Materials

In the present study four different conditions were compared, namely: 1. hands-on lessons, 2. online lessons, 3. online lesson and then hands-on lesson and 4. hand-on lesson and then online lesson. All the conditions consisted of two lessons. Condition 3 and 4 used the same lessons as condition 1 and 2 but the former only had one lesson of the two types of lessons instead of two lessons with the same type of material. For example, condition 3 started with the first online lesson that also condition 2 had and subsequently followed the second hands-on lesson as the same as condition 1 had.

The learning and teaching materials for all conditions were based on design principles that were derived from literature about misconceptions such as introducing anomalous data (Chinn and Brewer, 1993), introducing scientific conceptions that are intelligible, fruitful and plausible (Posner et al., 1982) and adopting inquiry-based learning as an educational strategy (e.g. Zacharia et al., 2015). The hands-on lesson and online lesson were designed to use the same type of knowledge (domain specific knowledge about Newton's first law) and type of instruction (inquiry learning) and differed only at the aspect of using physical vs. online material in order to have an appropriate comparison as recommended by Klahr et al. (2007). Inquiry-based learning has been characterised as an educational strategy in which students were adopting a scientific approach in order to discover and construct new knowledge (De Jong, 2006; Pedaste et al., 2015). The five main phases of inquiry learning (namely orientation, conceptualization, investigation, conclusion and discussion) as described by Pedaste et al. (2015) were used to guide the inquiry learning in the hands-on lesson as well as the online lesson. The first phase orientation is intended to stimulate interest and curiosity in students for a new topic to investigate. Concretely this means for the lessons of the present research that the lessons were started plenary by orienting on the topic of Newton's first law by asking question(s) to students about concrete examples. Secondly, the students started with the learning activities in pairs. Each learning activity went through the phases conceptualisation, investigation, conclusion and discussion. The conceptualization phase started with question(s) which incited predictions of the students. Subsequently in the investigation phase, students tested their predictions by experimenting with real hands-on material or online simulations. Next, students drew conclusions from the data and finally evaluated if they had expected their findings in the conclusion phase. After each learning activity, plenary discussions were lead by the teacher which contributed to the discussion phase. In the discussion phase students present and evaluate their outcomes of an inquiry phase or the whole process to others and collect feedback from them. Every lesson included two or three learning activities. At the end of a lesson, a final conclusion was drawn plenary. Thus the hands-on and online lesson used the same structure and discussed the same topic but differed on what kinds of materials were used for the experimentation phase.

The designed instruction materials were pre-tested. One primary school teacher and educational scientist, one physics teacher in training, and one physics teacher and researcher evaluated the instruction. Furthermore, three educational science students and one physics student offered feedback on the instruction materials. The learning materials were tested by 46 students from primary school as well as secondary school. As of result of the pre-testing, among other extra information about the resulting force was given, images were added and sentences rephrased.

Hands- on learning materials.

The hands-on lessons used a model of a roller coaster to explain Newton's first law. Students could experiment by moving a roller coaster cart on different roller coaster setups (e.g. hill and looping) and experimenting with masses on a cart. The theme of the lessons was roller coasters and started with questions from the teacher about examples of Newton's first law with regard to roller coasters. A worksheet guided the students through the phases of inquiry learning during the learning activities. Figure 1 shows the hands-on material of the first learning activity of the first lesson. In concrete terms this meant for the first learning activity that students predicted if the car continue moving as long on a track with fine sandpaper as on rough sandpaper or a on a track without sandpaper. Secondly, students released the cart on the different tracks and counted how many times the cart would move back and forth till the cart stopped. Those results and conclusions were written down on the worksheets and also their evaluation of their findings. Furthermore, a final conclusion at the end of the lesson was written and after each learning activity a plenary discussion was lead by the teacher.

Figure 1. Hands-on material

Online learning materials

The online learning environment was a go lab inquiry space about Newton's first law, divided in two lessons: (http://graasp.eu/ils/56b5f2a92aec72857324571f/?lang=nl and http://graasp.eu/ils/56b5ece72aec72857324571e/?lang=nl). Each lesson started with a movie that introduced Newton's first law. Subsequently, the learning activities followed the structure of inquiry learning, which were the same as the phases in the worksheet of the hands-on lesson. The different aspect was that the online learning environment addressed Newton's first law by allowing students to experiment with simulations. Figure 2 shows a screenshot of the first learning activity of the first lesson. For example, students had to predict what the influence of friction on a halfpipe would be for a skateboarder. Secondly, students released the skateboarder on the halfpipe in a simulation and counted how many times the cart would move back and forth till the skateboarder stopped. Students could change the friction on the halfpipe. Their findings, conclusions and evaluation could be filled in online. Moreover, after each topic a classroom discussion was lead by the teacher. Finally, the final conclusion showed a short movie that referred to the first movie and the teacher drew also a conclusion plenary.

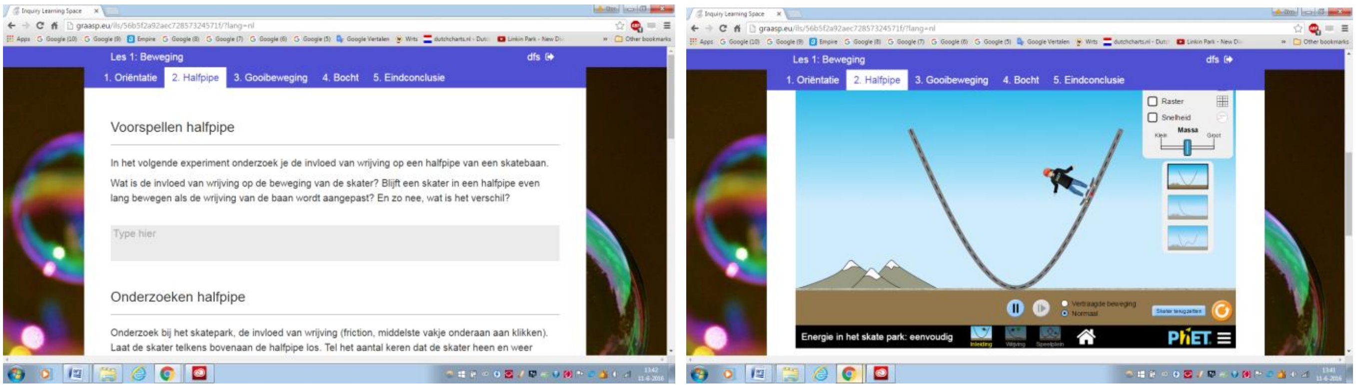


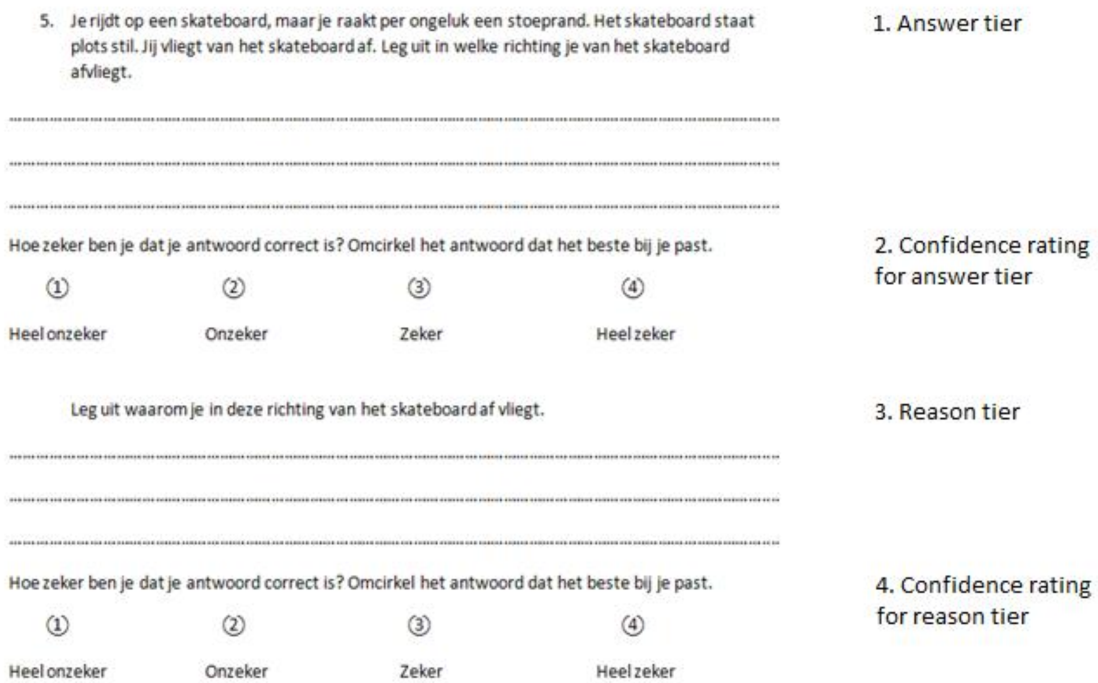
Figure 2. Online learning environment

Instruments

Test (mis)conceptions and confidence.

The questions of the (mis)conception and confidence test are based on the four-tier diagnostic test (e.g. Gurel et al., 2015; Yang & Lin, 2015). This type of test measures not only students' understanding of misconceptions but also students' confidence in their responses (Yang & Lin, 2015). The information about the level of confidence is valuable because it can help to determine the nature and strength of misconceptions as accurately as possible (Gurel et al., 2015; McClary & Bretz, 2012).

Misconceptions are very resistant to change and can impede further learning but even small shifts in conceptual change can be shown with a four-tier diagnostic test (Gurel et al., 2015). For example, starting to doubt about an own incorrect conception can be detected which is the first step of conceptual change according to Posner et al. (1982). Each task was build up as: 1. answer tier → 2. confidence rating for answer tier → 3. reason tier → 4. confidence rating for reason tier. The addition of reason tiers with corresponding confidence tiers helps to determine whether and how strong (mis)conceptions are biased by students reasoning (Caleon & Subramaniam, 2010; McClary & Bretz, 2012). For example, students could reason incorrectly but that would result in a correct answer or students could answer and reason correctly (Caleon & Subramaniam, 2010). Figure 3 displays an example of a (mis)conceptions test question. The first questions started with an answer tier which were open questions, multiple-choice questions or questions in which students were invited to draw pictures to demonstrate their knowledge (e.g. see "1. Answer tier" in Figure 3). Subsequently, students were asked how sure they were about their first answer (e.g. see "2. Confidence rating for answer tier" in Figure 3). The reason tier was an open question to invite students to write down why they chose their answer at the first question in order for them to show their line of reasoning (e.g. see "3. Reason tier"



5. Je rijdt op een skateboard, maar je raakt per ongeluk een stoeprand. Het skateboard staat plots stil. Jij vliegt van het skateboard af. Leg uit in welke richting je van het skateboard afvliegt.

1. Answer tier

Hoe zeker ben je dat je antwoord correct is? Omcirkel het antwoord dat het beste bij je past.

①	②	③	④
Heel onzeker	Onzeker	Zeker	Heel zeker

2. Confidence rating for answer tier

Leg uit waarom je in deze richting van het skateboard af vliegt.

3. Reason tier

Hoe zeker ben je dat je antwoord correct is? Omcirkel het antwoord dat het beste bij je past.

①	②	③	④
Heel onzeker	Onzeker	Zeker	Heel zeker

4. Confidence rating for reason tier

Figure 3. An example question of test(mis)conception and confidence

in Figure 3). Finally, students were asked how sure they were about their reasons for their answer (e.g. see "4. Confidence rating for reason tier" in Figure 3). The (mis)conception test consisted of 8 questions of which 5 questions used the structure of the four-tier diagnostic test. The other 3 questions consisted of only one open question and a confidence measure about that particular question. This structure was used because the structure of the four-tier diagnostic test was not possible due to for example asking only factual knowledge or arguments for a phenomenon.

The parallel tests for measuring the (mis)conception and confidence were pre-tested because it was a newly constructed test for the current research. One Physics student and 11 secondary school students completed the tests and they had the possibility to write remarks in the margin. Besides that, 3 Educational science students, 1 primary school teacher and educational scientist, 1 physics teacher in training and 1 physics teacher and researcher provided feedback about the test. As a result of the pre-testing, the number of questions was reduced, some sentences were reframed, important words highlighted and clearer images were used.

Cronbach's alpha was calculated for the test of the (mis)conception and confidence in order to ensure reliability (Peterson, 1994). Cronbach's Alpha was calculated for the (mis)conception tests and confidence, which were consequently for the pre-test conception $\alpha = .24$ and pre-test confidence $\alpha = .79$, for post-test 1 conception $\alpha = .62$ and post-test 1 confidence $\alpha = .91$ and finally post-test 2 conception $\alpha = .63$ and post-test 2 $\alpha = .92$. Those Cronbach's Alpha's would be classified as acceptable by the standards of Murphy and Davidshofer (1988) in Peterson (1994), except for the pre-test conception. However, this was probably due to student's sporadic or lack of knowledge because they were not yet taught about Newton's first law and as a result of that answered single questions correctly by guessing.

Science curiosity scale.

Students' interest in science was measured by the Science curiosity scale of Harty and Beall (1984). This scale was translated and adapted to the Dutch context. This scale consisted of 26 items on a five point likert scale (1= strongly disagree; 5= strongly agree). Score 1 and 2 were considered as a negative viewpoint on science, score 3 as neutral viewpoint on science and scores 4 and 5 as positive viewpoint on science. Example questions were: *‘Wetenschappelijke tijdschriften zijn interessant.’* and *‘Ik zou het leuk vinden om wetenschappers te horen praten over hun beroep.’*. The science curiosity scale was not pre-tested with the target audience because this was a frequently used instrument (e.g. Rubenstein, 2000, Sharp & Kuerbis, 2006). However, one Psychology and Educational science student did offer feedback. After deleting three items which did not correlate well with the overall science curiosity scale, the total scale of the science curiosity scale at the pre-test had $\alpha = .90$ and the total scale of the science curiosity scale at the post-test 2 had $\alpha = .92$. Those Cronbach’s Alpha’s were high according to the standards of Murphy and Davidshofer (1988) in Peterson (1994).

Procedure

First, the ethical committee of the University of Twente approved the current study and the execution of related research activities. The participating schools distributed an information letter to the students and parents, which included among others a passive informed consent and the goal of the research and research procedure. Parents that had objections could call or mail before the research started. The pre-test consisted of the science curiosity scale and a test that measured the (mis)conception and confidence. Based on the (mis)conception test results, students were assigned to the different conditions of the experiment. Approximately four weeks after the pre-test, the different lessons were taught by a physics teacher and the researcher. Directly after the lessons, post-test 1 was administrated which included the test that measured (mis)conception and confidence. Post-test 2 was conducted three weeks after the intervention, including the (mis)conception and confidence test and science curiosity scale. Directly after post-test 2, students were thanked and debriefed about the research. The collected data were analysed, after which conclusions were drawn. Figure 4 shows a schematic representation of the procedure with regard to the instruments.

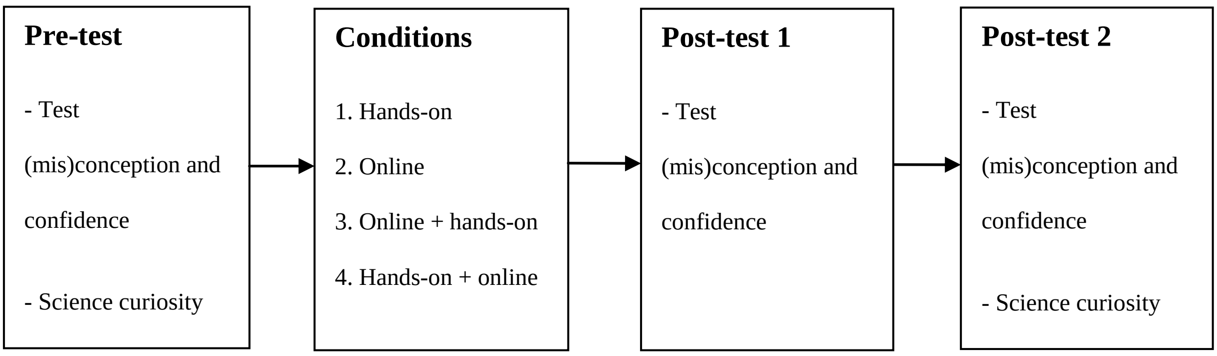


Figure 4. Overview procedure

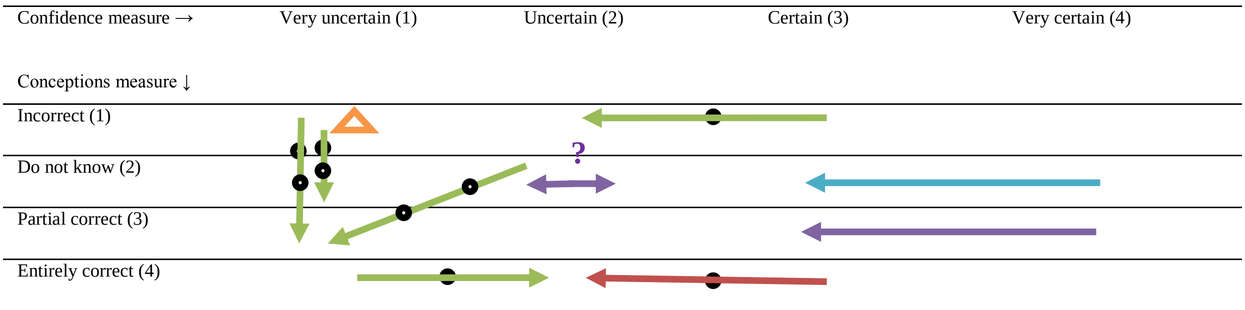
Coding

The (mis)conception and confidence test were coded in the following steps. The multiple-choice questions were scored '0' for an incorrect answer and '1' for a correct answer. The open questions were examined using a rubric. For the answer tier of the misconceptions test '0', '0,5'and '1'were scored for each incorrect, partial correct and correct response, respectively. The rubric varied for the reason tier from 0 points for no reasoning or misconception, 1 point for a partial correct answer and 2 points for a correct display of scientific conception. The minimum score for the (mis)conception test was 0 and the maximum score for the (mis)conception test was 20.

The scoring for the confidence ratings varied from 1, very uncertain till 4, very certain. When a student encircled two boxes, the lowest rating was coded because people have the tendency to present themselves in the best possible light (Randall & Fernandes, 1991).

Table 2

Overview types of shifts



Note. A green arrow is a positive shift; A red arrow is a negative shift; A blue arrow is an irrelevant shift; A purple arrow is an unknown shift; The orange triangle means no shift; One black circle means a small shift; Two black circles mean a large shift.

For the analysis of the confidence measures, the responses on the answer tier and reason tier were recoded. The scores were '1' for an incorrect answer, '2' when a student did not know the answer (e.g. "gamble", "-", "?"), '3' for a partial correct answer and '4' for an entirely correct answer. Subsequently, type of shift were coded between different tests (pair pre-test – post-test 1, pair pre-test – post-test 2, pair post-test 1- post-test 2). The types of shifts possible are indicated in Table 2. A positive shift was coded when a student attained more knowledge, would be more sure about their entirely correct answer or would be more unsure about their own incorrect answer (green arrow in Table 2). A negative shift would be if the opposite occurred than explained for a positive shift (red arrow in Table 2). An irrelevant shift would be when a student did not know and also later did not know the answer (blue arrow in Table 2). Besides that, a shift in rating of the confidence measure when a student had a partial correct answer is an unknown shift because it is not explicitly clear from the data if a student started to have more doubts about the correct part of the answer or the incorrect part. Furthermore, if a student stayed at the same level of conception and one or both confidence measures are missing, it was also coded as an unknown shift. If a student scored the same at both tests, than it was assigned as no shift (orange triangle in Table 2). Furthermore, there were also scores assigned for the size of the shift. When there is a shift in a conception, than it would be assigned as a large shift (two circles in an arrow indicates a large shift in Table 2). When a shift is only in confidence but the conception is at the same level, than it would be assigned as a small shift (one circle in an arrow indicates a small shift in Table 2).

A researcher, who was not involved in the current research project, served to establish a measure of inter-rater reliability by checking the tests in order to prevent bias of the researcher and to ensure quality. The Cohen's Kappa of the (mis)conception tests were 0.928, which was deemed as very high (McHugh, 2012). The Cohen's Kappa was not calculated for the confidence measures because it was self-evident how to score as the lowest encircled box directly represents a rating.

Results

If the conditions were indeed comparable based on gender, performance and science curiosity at the pre-test, chi-square tests were performed. The Chi-square test showed that the conditions did not significantly differ in gender (χ^2 (3, N = 127) = .53, p= .911) and performance on the pre-test (χ^2 (27, N = 127) =20.35, p=.816). Furthermore, students did not significantly differ in viewpoints on science curiosity (χ^2*negative viewpoint* (57, N = 109) =53.96, p= .590; χ^2*neutral viewpoint* (39, N = 109) = 43.39, p= .290; χ^2*postive viewpoint* (66, N = 109) = 60.41, p= .671). That was the reason why further analyses were not corrected by gender, science curiosity or performance on pre-test.

Misconceptions test

In order to determine whether there were statistically significant differences between the four different conditions for tackling students' misconceptions about Newton's first law, repeated-measures ANOVA were executed. A difference was considered as significant when the p-value was smaller than .05. Repeated-measures ANOVA were performed with between-subjects variable, the condition in which each student participated, and within-subject variable, repeated measures of the total performance scores on the misconceptions tests. Mauchly's test indicated that the assumption of sphericity has been met, χ^2 (2)= .41, *p*=.817. In other words, the results showed that the total performance scores were not significantly affected by the condition. Means and standard deviations of performance scores are shown in Table 3. Participants in the hands-on condition seemed to have the most improved scores compared to the other conditions but there was as said not a significant difference between conditions (see Table 3).

Table 3

Mean values and standard deviation of performance scores

Condition	Pre-test		Post-test 1		Post-test 2	
	Mean	SD	Mean	SD	Mean	SD
Hands-on (n=32)	3.39	1.85	7.36	3.32	7.28	2.81
Online (n=34)	3.41	1.50	6.22	2.26	7.07	2.96
Online + Hands-on (n=31)	3.29	1.49	6.87	2.35	6.35	2.30
Hands-on + online (n=30)	3.40	1.50	5.95	2.38	6.53	2.33
Total (n=127)	3.37	1.57	6.60	2.64	6.82	2.63

Repeated measures showed a main effect for the performance scores with regard to test moments, *F* (2)= 134.17, *p*=.000. The Bonferonni test showed as post-hoc analysis that the students significantly improved their scores on the conception test comparing post-test 1 with the pre-test (*M* = 3.23, *SD* = .24, *p*=.000) as well as post-test 2 with the pre-test (*M* = 3.44, *SD* = .23, *p*=.000). However, there was no significant increase between post-test 2 comparing with post-test 1 (*M* = .21, *SD* = .23, *p*=1.000), which indicated that student's knowledge about Newton's first law did not significantly changed after three weeks. The learning gains of post-test 1 compared to the pre-test for the four conditions are shown in Figure 5.

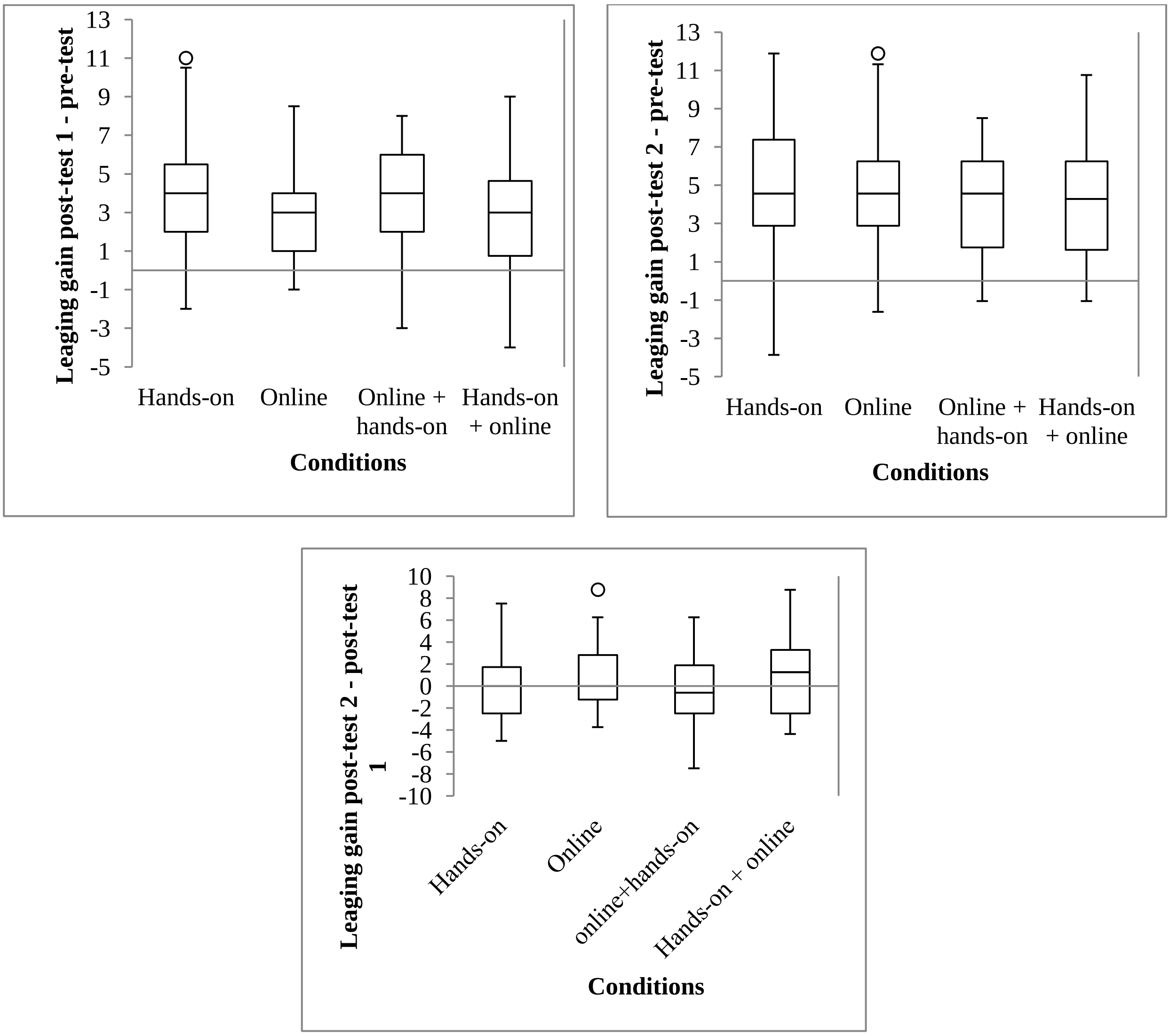


Figure 5. Learning gain scores between different tests.

Confidence rating

The aim of the confidence ratings was to measure changes in confidence and determine shifts in conceptual change in nature and strength. The mean values and standard deviations of mean total confidence scores of the participants about their answers at the performance tests are presented in Table 4. Repeated-measures ANOVA were performed with between-subjects variable, the condition in which each student participated, and within-subject variable, repeated measures of the mean confidence scores on the misconceptions tests. Mauchly's test indicated that the assumption of sphericity has been met (χ^2 (2)= 3.18, *p*=.204). The participants were fairly confident about their answers on the tests, in all conditions ($M_{total\ pre\text{-}test}$ = 2.62; $M_{total\ post\text{-}test\ 1}$ = 2.99; $M_{total\ post\text{-}test\ 2}$ = 2.86).

Table 4.

Mean values and standard devations of mean total confidence scores on the (mis)conceptiontests

Condition	Pre-test		Post-test 1		Post-test 2	
	Mean	SD	Mean	SD	Mean	SD
Hands-on	2.58	.48	3.02	.60	2.87	.60
Online	2.62	.50	2.94	.59	2.80	.61
Online + Hands-on	2.64	.40	2.97	.47	2.85	.44
Hands-on + online	2.65	.45	3.02	.54	2.91	.64
Total	2.62	.45	2.99	.55	2.86	.58

Note. N=127 However, not all students answered at all confidence questions. Mean values and standard deviations are computed of known confidence ratings

Repeated measures showed a main effect for the confidence scores with regard to test moments (*F*(2)= 51.80, *p*=.000). Post hoc analysis showed that the students' mean confidence ratings increased significantly directly after the lessons compared to the pre-test (*M*= .37, *SD*= .04, *p* = .000 and three weeks after the lessons compared to the pre-test (*M*=.24 , *SD*= .04, *p* =.000). Between mean confidence scores on post-test 2 and mean confidence scores on post-test 1 were, *M* = -.13, *SD*= .04, *p* = .004. Thus between post-test 2 and post-test 1 was a significant negative relationship with regard to mean confidence ratings.

Table 5 report the mean values of correctness for both the answer and the reason tiers, their corresponding mean confidence rating, and t-test values between confidence ratings in the answer tier and reason tier. Three questions were not included in these analyses because those questions did not have the structure answer tier- answer tier confidence ratings – reason tier- reason tier confidence ratings. 127 Students participated in the analysis of the (mis)conception tests. However, not all students answered all confidence questions. Mean values and standard deviations were computed of known confidence ratings. Furthermore, none of the students answered the answer rating that they did not know the answer to which explains mean values of 0 and no standard deviations or paired sample t-test in Table 5. Additionally, in Table 5 no standard deviation was mentioned for a partial correct

Table 5

Mean values and standard deviations of correctness and confidence ratings.

	ANS	(SD)	ACÖNF	(SD)	RSN	(SD)	RCONF	(SD)	ACONF-RCONF
Pre-test									
- Incorrect	.57	.20	2.43	.54	.87	.17	2.68	.57	.000*
- Do not know	.00	.00	-	-.	.06	.11	1.97	1.11	-
- Partial correct	.00	.02.	4.00	-	.06	.12	3.13	.61	-
- Entirely correct	.43	.20	2.99	.70	.01	.04	2.60	1.14	.354
Post-test1									
- Incorrect	.38	.23	2.53	1.19	.62	.23	2.81	.76	.004*
- Do not know	.00	.00	-	-	.14	.11	.24	.79	-
- Partial correct	.05	.11	.61	1.23	.28	.23	2.35	1.49	.000*
- Entirely correct	.57	.23	3.30	.63	.07	.12	1.01	1.56	.000*
Post-test 2									
- Incorrect	.29	.20	2.26	1.29	.58	.26	2.64	.97	.005*
- Do not know	.00	.00	-	-	.03	.09	.15	.44	-
- Partial correct	.06	.10	2.67	.96	.33	.24	3.04	.64	.130
- Entirely correct	.65	.21	3.09	.60	.06	.10	3.15	.74	.497

Notes: ANS = Answer tier; ACONF = Answer tier confidence ratings; RSN = Reason tier; RCONF = Reason tier confidence ratings; ACONF-RCONF = t-test comparison between answer tier confidence rating and reason tier confidence ratings.; Incorrect = misconception or no answer; Do not know = student showed in their answer that they did not know the answer; Partial correct = student showed some conception but is not entirely correct; Entirely correct = completely correct conception.
* $p < .05$

answer at the pre-test because only one participant had that score and no paired sample t-test score was mentioned because there were no valid pairs. Repeated measures showed no significant effect between conditions for all types of correctness and confidence tiers.

A paired-sample t-test was conducted to compare the answer confidence rating with the reason confidence rating at each test moment. The paired-sample t-test for the incorrect answers at the pre-test showed a significant negative relationship between the confidence rating on the answer tier and the confidence rating on the reason tier at the pre-test ($t(125)=-5.34$, $p=.000$). These results suggest that when students answered incorrectly, they were more confident about the reason tier. Moreover, students were more often incorrect about their reasons than their answers ($M_{ANS\ incorrect}=.57$, $M_{RSN\ incorrect}=.87$). Besides that, no significant positive relationship was found between de the confidence scores on the answer tier and reason tier if students were correct at the pre-test ($t(4)=1.05$, $p=.354$).

For post-test 1 paired sample t-test showed that students were significantly more sure about their reason tier than their answer tier if they answered incorrectly and did not know the answer ($t_{incorrect}(126)$ -2.91, $p_{incorrect}$ =.004; $t_{do\ not\ know}(126)$= -3.47, $p_{do\ not\ know}$ =.001). Students were more incorrect about their reason tier than the answer tier ($M_{ANS\ incorrect}$=.38, $M_{RSN\ incorrect}$=.62). On the other hand, if students were entirely correct, they were clearly more sure about their answer tier than their reason tier ($t_{entirely\ correct}(126)$= 15.65, $p_{entirely\ correct}$ =.000).

For post-test 2 paired sample t-test indicated that students were significantly more sure about their reason tier than their answer tier if they answered incorrectly ($t_{incorrect}(124)$=-2.86, $p_{incorrect}$ =.005). Students were again more incorrect about their reason tiers than answer tiers ($M_{ANSincorrect}$=.29, $M_{RSN\ incorrect}$=.58). Nonetheless, it showed no other significant relationships between the answer confidence tier and reason confidence tiers ($t_{partial\ correct}$ (27)= -1.56, $p_{partial\ correct}$ =.130; $t_{entirely\ correct}(30)$= 69, $p_{entirely\ correct}$ =.497) .

Table 6 shows the mean values (percentages) and standard deviations for the proportion of the answer confidence tier and reason confidence tier. When the students' mean confidence scores on the answer tier and the confidence score on the reason tier were compared, it was found that the former was significantly higher than the latter at post-test 1 and post-test 2 ($t_{post\text{-}test1}(118)$=5.50, p=.000; $t_{post\text{-}test2}(113)$=4,03 p=.000). There was no significant difference between the confidence rating on the answer tier and the reason tier at the pre-test($t(119)$=-.32, p=.753). Furthermore, the answer confidence ratings and reason confidence rating remained the same, had in percentages the highest proportion

Table 6

Mean values (percentages) and standard deviations for proportion answer confidence tier and reason confidence tier

Condition	Pre-test		Post-test 1		Post-test 2	
	Mean	SD	Mean	SD	Mean	SD
ACONF = RCONF (%)	62.68	24.86	68.35	24.62	65.51	24.65
ACONF > RCONF (%)	18.43	18.45	22.05	20.44	21.26	20.43
ACONF < RCONF (%)	17.64	19.46	8.19	12.69	10.71	15.90

Note. N=127 However, not all students answered at all confidence questions. That is why the the total of the mean values (percentages) is not 100%.

($M_{pre\text{-}test}$= 62.68, $SD_{pre\text{-}test}$ = 24.86; $M_{post\text{-}test\ 1}$= 68.35, $SD_{post\text{-}test\ 1}$ =24.62 $M_{post\text{-}test\ 2}$= 65.51, $SD_{post\text{-}test\ 2}$ = 24.65). Subsequently, the answer confidence rating was higher than the reason confidence rating and least present was answer confidence rating was smaller than the reason confidence rating. Repeated measures showed no significant difference between the four different conditions with regard to proportion of the answer confidence tier and reason confidence tier ($F_{ACONF=RCONF}(5.64)=1.51$, p=.181; $F_{ACONF>RCONF}(6)=.53$, p=.786; $F_{ACONF<RCONF}(5.18)=1.51$, p=.185).

Furthermore, the types of shifts were coded (see Table 7). One-way ANOVA showed no significant differences between the four different conditions with regard to type of shifts. Except for no shift between post-test 1 and post-test ($F(3,123)=3.02$, p=.032), for which Tukey's HSD post hoc test showed that the hands-on condition had more no shifts than the hands-on and online condition (p=.030). Comparing the post-tests to the pre-test, mainly positive shifts occurred and those types of positive shifts were mainly large positive shifts (see Table 7). Subsequently, negative shifts, which were mainly small shifts, occurred the most after positive shifts when comparing the pre-test to the post-tests (see Table 7). For post-test 1 – post-test 2, negative shifts were found most frequently and also mostly large negative shifts rather than small negative shifts (see Table 7). In order to check whether there were significant differences in shifts between the different shifts moments, chi-square

Table 7

Mean values (percentages) and standarddeviations for types of shifts by condition

Condition	Pre-test – Post-test 1		Pre-test - Post-test 2		Post-test 1-Post-test 2	
	Mean	SD	Mean	SD	Mean	SD
Positive shift	46.15	15.38	47.30	15.33	31.01	14.89
- Small positive shift	13.08	10.13	13.33	10.00	10.42	9.27
-Large positive shift	33.07	14.64	33.98	14.62	20.47	12.31
Negative shift	27.14	13.56	26.17	14.48	31.80	14.06
- Small negative shift	17.02	12.09	16.72	12.07	13.14	10.84
- Large negative shift	9.93	8.73	9.45	8.62	18.66	11.36
Irrelevant shift	.06	.68	.00	.00	.06	.68
Unknown shift	2.91	5.03	3.27	6.09	6.90	7.04
No shift	23.74	12.03	23.26	14.26	30.22	16.11

Note. N=127

tests were performed. Chi-square tests indicated that between the shift of pre-test to post-test 1 and post-test 1 to post-test 2 significant difference for large negative shifts (χ^2 (30) = 45.74, *p*=.033), thus between post- test 1 and post-test 2 more large negative shifts occurred than between the pre-test and post-test 1. For all test moments shifts, no shifts occurred as the most and unknown and irrelevant shifts were the rarest (see Table 7). Chi-square tests showed that the differences between all shifts moments were significant for unknown shifts ($\chi^2_{\textit{Pre-test – Post-test 1} \rightarrow \text{Pre-test - Post-test 2}}$ (12) = 51.89, *p*=.000; $\chi^2_{\textit{Pre-test – Post-test 1} \rightarrow \text{Post-test 1- Post-test 2}}$ (12) = 33.71, *p*=.001; $\chi^2_{\textit{Pre-test – Post-test 2} \rightarrow \text{Post-test 1- Post-test 2}}$ (16) = 70.18, *p*=.000), which indicated that after each analysis more data is unknown. Furthermore, the shifts of pre-test to post-test 1 between pre-test to post-test 2 were significant for small positive shifts (χ^2 (36) = 59.00, *p*=.009, indicating more shifts) large positive shifts (χ^2 (81) = 152.28, *p*=.000, indicating more shifts), large negative shifts (χ^2 (25) = 80.83, *p*=.000, indicating more shifts) and no shifts (χ^2 (63) = 88.23, *p*=.020, indicating less shifts).

Science curiosity

Students' interest in science was also taken into account. The mean values percentages and standard deviations of viewpoints on science at the pre-test and the gain scores are shown in Table 8. Gain scores-between the pre-test and post-test 2 were calculated by subtracting the post-test 2 score from the pre-test score. The data showed that positive viewpoints on science were reported most frequently across all conditions and measurement moments (see Table 8). Furthermore, negative viewpoints were after positive viewpoints tended to report as more frequently, in particular at the post-test 2 (see Table 8). In order to determine whether there were statistically significant differences between conditions, one-way ANOVA were performed. The students in the online and then hands-on condition were the only students that often indicated positive viewpoints on science compared to the pre-test, in contrary to the other conditions, but this difference was not statistically significant (*F*(3, 101)=.81, *p*=.492). There was also no statistically significant difference between conditions for neutral viewpoints on science (*F*(3, 101)=.90, *p*=.446) or for negative viewpoints on science (*F*(3, 101) = 1.31, *p*=.276). Moreover, paired sample t-test showed no significant differences for the different viewpoints ($t_{negative\ viewpoint}(104) = 1.71$, $p_{negative\ viewpoint} = .091$; $t_{neutral\ viewpoint}(104) = -.03$, $p_{neutral\ viewpoint} = .977$; $t_{positive\ viewpoint}(104) = -1.38$, $p_{positive\ viewpoint} = .171$).

Table 8

Mean values (percentages) and standard deviations of viewpoints on science

Condition	Pre-test						Gain scores					
	Negative		Neutral		Positive		Negative		Neutral		Positive	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Hands-on (*n*=26/31[a])	28.60	26.47	24.75	15.88	46.66	23.80	-3.45	15.47	0.70	12.59	-0.35	18.05
Online (*n*=28/32 [a])	25.47	18.55	25.93	11.74	47.36	22.34	2.90	14.72	3.22	17.58	-4.83	13.82
Online + Hands-on (*n*=29/29 [a])	31.63	21.80	25.64	13.52	42.73	21.40	.32	14.62	-1.13	13.78	0.81	15.25
Hands-on + online (*n*=26/30 [a])	27.42	21.08	28.09	10.45	44.48	20.22	7.02	14.66	-3.01	13.17	-4.01	15.74
Total (*n*=109/122 [a])	28.32	21.90	26.09	12.92	45.27	21.74	2.48	14.93	-0.04	14.44	-2.11	15.70

Note. [a] The first and second figure represent respectively the number of participants at the pre-test and post-test. The number of participants differed on the the pre-test and post-test 2 of the science curiosity scale because only at the misconceptiontest all the missing data was removed. That is why also the paired sample t-test of the science curiosity scale had a different number of participants.

Conclusion and discussion

The current research inquired what the most effective method is, comparing hands-on lessons, online lessons or combinations of both (online lesson and then hands-on lesson; hand-on lesson and then online lesson), for tackling students' misconceptions about Newton's first law. This study showed that none of the type of methods could be regarded as the most effective method for tackling students' misconceptions about Newton's first law, even taken into account the secondary questions as performances on (mis)conception test, long-term recall, students' confidence in their (mis)conception and students' interest in science.

This research contributes to the scientific debate whether hands-on lessons or online lessons would have more affordances with respect to tackling students' misconceptions and whether a combination of both would be more beneficial, and if so, if the order of presentation would matter (Klahr et al., 2007; Lazonder & Ehrenhard, 2014; Zacharia, 2007). It appeared from the present study that no method or combination of methods was superior to another method in order to tackle students' misconceptions about Newton's first law. Students have gained comparable conception in all conditions and did not score significantly worse or better after three weeks than directly after the lessons. The findings of the current study are therefore in contradiction with several studies that favour hands-on lessons (e.g Zacharia, et al., 2012), online lessons (e.g. Finkelstein et al., 2005) or a combination of both as being better than one of its constituents (e.g. Chini et al., 2012; Smith & Puntambekar, 2010; Zacharia, 2007), since the current study showed that not one of the types of methods was more beneficial than the others. The present study is more in line with studies that show that hands-on lessons and online lessons are equally effective in achieving several instructional objectives (e.g. Triona &Klahr, 2003), or that the four conditions as used in the current study are equally effective in promoting conceptual change (Zacharia & Olympiou, 2011). Zacharia and Olympiou (2011) explained those findings by stating that experimentation, either hands-on or online experimentation, and not physicality is important in their domain of physics learning. Triona and Klahr (2003) suggested the idea that the type of material does not make the difference as long the instruction method is held constant. Triona and Klahr (2003) said that the most important features were captured in the instruction thus the type of material did not matter in their learning context. Other explanations with regard to the current research could be that not all students from the sample were included in the analysis. Furthermore, the first two classes had some unexpected events and a different design that could have influenced the results, such as some students and teachers were surprised that the research took place at the specific date and two classes had the lessons at the same time instead of only class. Additionally, for all the classes the hands-on lessons were taught by the researcher instead of a physics teacher. Taking into account the important role of teachers for the students’ achievements, this could influence the results because the researcher was not an experienced teacher (e.g. Hattie, 2003; George & Kaplan, 1998). The absence of decrease in performance after three weeks could be viewed as surprising because memory systematically deteriorates with delay (Anderson, 2005). An explanation for the lack of loss with delay could be that Newton’s first law and inertia was also a topic in the curriculum at one location during the period of the research. Students had an exam about forces the week before post-test 2. So this could explain why students still had quite good retention scores, since they recently rehearsed and studied the topic. Student’s practice would diminish loss with delay according to Anderson (2005).

Furthermore, the current research contributes to the knowledge base about confidence and misconceptions. No method or combination of methods stood out with regard to how sure students were about their answers. Students appeared to be more confident about their answers directly after the lessons but were less confident three weeks later than directly after the lessons. Yang and Lin (2015) drew the conclusion that students' mean confidence rating for the answer tier was significantly higher than for the reason tier. This is in line with the current study that indicated that overall, students tended to be more confident about their answering tier than their reasoning tier. However, when only incorrect answers were compared in the current study, students were more sure about the reasoning tier instead. An explanation for the higher mean rating on the answering tier than the reasoning tier could be that students were more often correct about their answering tier than reasoning tier and students scored more often higher on the confidence rating when they were entirely correct than incorrect. Besides that, previous studies using the four-tier diagnostic usually had only one moment of measurement (e.g. Caleon & Subramaniam, 2010; Kaltakçi, 2012; Yang & Lin, 2015), but the present study had three moments with the same sample, which as a result could identify types of shifts. For example, between the pre-test and post-test 1 were more small and large positive shifts and large negative shifts present than between the pre-test and post-test 2.

Moreover, the present study is also valuable for science as a methodological development for measuring conceptual change. Gurel et al. (2015) pointed out that all diagnostic assessment methods in their review had their own strengths and limitations but also mentioned that more emphasis should be given to four tier tests in all fields of science. Measuring confidence on the answer and reasoning tier also offered opportunities to assess the nature and strength of misconceptions (Gurel et al., 2015). The four tier tests were often multiple-choice tests that measure confidence ratings for an answer tier and a reasoning tier (Gurel et al., 2015; Yang & Lin, 2015). Nonetheless, in the current research mainly open questions were asked. Which as result made it possible to more exactly determine if a student would have known the correct answer or was just guessing and offered the possibility to construct and demonstrate their own answers instead of selecting from a shortlist (Kaltakçi, 2012). It is important to determine as correctly as possible conception shifts because misconceptions are very resistant to change and now also small shifts or even negative shifts that otherwise would not be noticed in conceptual change could be measured (e.g. being more unsure about a correct answer after the lessons). The tests also appeared to be reliable, except for the pre-test. The Cronbach’s Alpha of the first (mis)conception test was very low. This could be due to student’s lack of knowledge or sporadic knowledge and as a result of that answered some single questions correctly by guessing. However, a few students may had have a little bit of knowledge because at one school the term inertia was shortly mentioned in a movie before the pre-test.

Furthermore, the present study is significant for society because students’ viewpoints on science impact their future career choices and there is growing need for STEM jobs in modern society (Jocz et al., 2014; Morais, 2015; Wysocki et al., 2013). Several initiatives promote science curiosity such as FabLab+ and Miss to Hit. Those initiatives support teachers with a range of lessons in order to incite students to science and technology. Those initiatives can implement findings of the present study in their projects because this study also researched whether an instruction method would increase the science curiosity of students. The current study did not show any preferred instruction method in order to increase science curiosity. The findings even showed that science curiosity did not significantly increase. Literature also suggest that opinions about science are founded early in childhood education and are hard to change later in children’s educational career (Kermani & Aldemir, 2015; Morais, 2015). Furthermore, even more students lose their interest in science during secondary school (Krapp & Prenzel, 2011; Logan, & Skamp, 2013). Other factors than the lessons could also have played a role concerning the lack of increase in students’ science curiosity in the present study. Students had to fill in the science curiosity questionnaire directly after an exam week or two weeks holiday, which could influence their decreased motivation for school and science. Furthermore, during the same period of the research, students had to make their final decision for their further study specialisation. If students had already made the decision to stop with the science subjects at the moment when this research took place, it would probably result in lower scores on the science curiosity scale. The findings now arouse the suggestion that only one or two lessons could not change students’ science curiosity positively. Possibly more long-term lessons/experiments and science integrated in the curriculum and connected with a daily relevant context could increase more science curiosity (Krapp & Prenzel, 2011; Logan, & Skamp, 2013). Future research can focus on the question if those kinds of projects are able to increase students’ science curiosity.

The current study is of value for educational practise as well because it is informative about which instruction method, a hands-on lesson, an online lesson or a combination of both, is preferable to teach Newton's first law. Currently this information is missing in a framework for science education (Zacharia, 2008; Zacharia et al., 2012). The present study indicated that no instruction method was more beneficial than another. In a framework for science education, no directive guidelines would be necessary to include which method teachers should chose. Teachers should have the option to choose their own preferable teaching method, since it does not seem to matter from which method students would learn more or make them more enthusiastic about science. Although it can be pointed out to teachers what the benefits are of the different methods, like an online lesson will be safer, cheaper, less time consuming and affords easier classroom management (Klahr et al., 2007; Smith & Puntambekar, 2010; Triona & Klahr, 2003; Zacharia et al., 2008). It is important to note that the current study has focused on students' conceptual understanding, confidence and science curiosity and not on students and teachers experiences of the lessons. Further research can also include those affective factors. Especially because all the methods were equally effective, the opinions of students and teachers could be a decisive factor on which type of method would be recommendable in a framework for the science curriculum. Furthermore, future research can make a fully blended learning lesson instead of a separate hands-on lesson and an online lesson that would represent the combination of two types of methods. For example, students can record on a camera how a roller coaster car can be launched from a ramp. Students can edit the movie online by drawing arrows and playing in slow motion. A full integration of the hands-on method and online method could be compared with only one type of method or combination of two separate lessons.

In conclusion, the present study showed that hands-on lessons, online lessons or combinations of both were equally effective for acquiring conceptual change and changing science curiosity. It will be worthwhile to expand the domains and different contexts, by using existing instruments like the four-tier tests or using other instruments like interviews, to add more knowledge for the science field of conceptual change as well as practical guidelines for practitioners. But for now, practitioners should have the chance to choose their own preference of instruction material to tackle misconceptions of students about Newton's first law.

References

Acar, Ö. (2014). Scientific reasoning, conceptual knowledge, & achievement differences between prospective science teachers having a consistent misconception and those having a scientific conception in an argumentation-based guided inquiry course. *Learning and Individual Differences, 30*, 148-154. doi:10.1016/j.lindif.2013.12.002

Anderson, J. R. (2005). *Cognitive psychology and its implications* (6th ed.). New York: Worth Publishers.

Boyes, E. (1988). Catastrophic misconceptions in science education. *Physics Education*, *23*(2), 105-109.

Bulunuz, N., & Jarrett, O. S. (2010). The effects of hands-on learning stations on building American elementary teachers' understanding about earth and space science concepts. *Eurasia Journal of Mathematics, Science & Technology Education*, *6*(2), 85-99.

Bulunuz, M., Jarrett, O. S., & Bulunuz, N. (2009). Middle school students' conceptions on physical properties of air. *Journal of Turkish Science Education*, *6*(1), 37-49.

Bursal, M. (2012). Changes in American preservice elementary teachers' efficacy beliefs and anxieties during a science methods course. *Science Education International, 23*(1), 40-55.

Caleon, I. S., & Subramaniam, R. (2010). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education, 40*(3), 313-337. doi:10.1007/s11165-009-9122-4

Carrejo, D. J., & Reinhartz, J. (2014). Facilitating conceptual change through modeling in the middle school science classroom. *Middle School Journal*, *46*(2), 10-19.

Chini, J. J., Madsen, A., Gire, E., Rebello, N. S., & Puntambekar, S. (2012). Exploration of factors that affect the comparative effectiveness of physical and virtual manipulatives in an undergraduate laboratory. *Physical Review Special Topics-Physics Education Research*, *8*(1), 010113. doi:10.1103/PhysRevSTPER.8.010113

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of educational research*, *63*(1), 1-49.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of physics*, *50*(1), 66-71.

De Jong, T. (2006). Technological Advances in Inquiry Learning. *Science, 312*, 532-533.

De Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and virtual laboratories in science and engineering education. *Science*, *340*(6130), 305-308.

DiGisi, L. L., & Yore, L. D. (1992, March). *Reading comprehension and metacognition in science: status, potential and future direction.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, Boston.

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

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., ... & LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical Review Special Topics - Physics Education Research*, *1*(1), 010103-1 - 010103-8. doi:10.1103/PhysRevSTPER.1.010103

Galili, I. & Tseitlin, M. (2003). Newton's first law: Text, translations, interpretations and physics education. *Science & Education, 12*, 45-73.

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

George, R., & Kaplan, D. (1998). A structural model of parent and teacher influences on science attitudes of eighth graders: Evidence from NELS: 88.*Science Education*, *82*(1), 93-109.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive psychology*, *12*(3), 306-355.

Gire, E., Carmichael, A., Chini, J. J., Rouinfar, A., Rebello, S., Smith, G., & Puntambekar, S. (2010). The effects of physical and virtual manipulatives on students' conceptual learning about pulleys. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th international conference of the learning sciences* (1, pp. 937 – 944). Chicago: International Society of the Learning Sciences.

Gurel, D. K., Eryılmaz, A. & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(5), 989-1008. Doi: 10.12973/eurasia.2015.1369a

Hadzigeorgiou, Y. (2001). The Role of Wonder and 'Romance' in Early Childhood Science Education Le Rôle de l'Émerveillement et du 'Roman' dans l'Éducation Scientifique des Jeunes Enfants El Papel de las Maravillas y 'Romance' en la Ciencia Educación Temprana de los Ninös. *International Journal of Early Years Education*, *9*(1), 63-69. doi: 10.1080/713670677

Harty, H., & Beall, D. (1984). Toward the development of a children's science curiosity measure. *Journal of Research in Science Teaching*, *21*(4), 425-436.

Hattie, J. (2003, October). *Teachers Make a Difference, What is the research evidence?* Paper presented at the Building Teacher Quality: What does the research tell us ACER Research Conference, Melbourne, Australia.

Heard, P. F., Divall, S. A., & Johnson, S. D. (2000). Can 'ears-on' help hands-on science learning for girls and boys?. *International Journal of Science Education*, *22*(11), 1133-1146. doi:10.1080/09500690050166715

Inan, H. Z., & Inan, T. (2015). 3 H s Education: Examining hands-on, heads-on and hearts-on early childhood science education. *International Journal of Science Education*, *37*(12), 1974-1991. doi: 10.1080/09500693.2015.1060369

Jocz, J.A., Zhai, J. & Tan, A.L. (2014). Inquiry learning in the Singaporean context: Factors affecting student interest in school science. *International Journal of Science Education, 36*(15), 2596-2618, doi:10.1080/09500693.2014.908327

Kaltakçi, D. (2012). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics (Unpublished doctoral thesis). Middle east technical university, Ankara, Turkey.

Kermani, H., & Aldemir, J. (2015). Preparing children for success: integrating science, math, and technology in early childhood classroom. *Early Child Development and Care*, *185*(9), 1504-1527. doi: 10.1080/03004430.2015.1007371

King, C. J. H. (2010). An analysis of misconceptions in science textbooks: Earth science in England and Wales. *International Journal of Science Education*, *32*(5), 565-601. doi:10.1080/09500690902721681

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of Research in Science teaching*,*44*(1), 183-203. doi: 10.1002/tea.20152

Kollöffel, B., & De Jong, T. (2013). Conceptual understanding of electrical circuits in secondary vocational engineering education: Combining traditional instruction with inquiry learning in a virtual lab. *Journal of engineering education*, *102*(3), 375-393.

Kozhevnikov, M., Gurlitt, J., & Kozhevnikov, M. (2013). Learning relative motion concepts in immersive and non-immersive virtual environments. *Journal of Science Education and Technology*, *22*(6), 952-962. doi: 10.1007/s10956-013-9441-0

Krapp, A., & Prenzel, M. (2011). Research on interest in science: Theories, methods, and findings. *International Journal of Science Education*, *33*(1), 27-50. doi:10.1080/09500693.2010.518645

Lazonder, A. W., & Ehrenhard, S. (2014). Relative effectiveness of physical and virtual manipulatives for conceptual change in science: how falling objects fall. *Journal of computer assisted learning*, *30*(2), 110-120. doi: 10.1111/jcal.12024

Levin, I., & Druyan, S. (1993). When sociocognitive transaction among peers fails: The case of misconceptions in science. *Child Development*, *64*(5), 1571-1591.

Levitt, K. E. (2001). An analysis of elementary teachers' beliefs regarding the teaching and learning of science. *Science education, 86*(1), 1-22. doi: 10.1002/sce.1042

Logan, M. R., & Skamp, K. R. (2013). The impact of teachers and their science teaching on students''science interest': A four-year study. *International Journal of Science Education*, *35*(17), 2879-2904. doi: 10.1080/09500693.2012.667167

Mayer, R.E. (2008). *Learning and Instruction* (2nd ed). Upper Saddle River, New Jersey: Pearson Merril Prentice Hall.

McClary, L. M., & Bretz, S. L. (2012). Development and assessment of a diagnostic tool to identify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education*, *34*(15), 2317-2341. doi: 10.1080/09500693.2012.684433

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122-130.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Mildenhall, P. T., & Williams, J. S. (2001). Instability in students' use of intuitive and Newtonian models to predict motion: the critical effect of the parameters involved. *International Journal of Science Education*, *23*(6), 643-660. doi: 10.1080/09500690117839

Morais, C. (2015). Storytelling with chemistry and related hands-on activities: Informal learning experiences to prevent "chemophobia" and promote young children's scientific literacy. *Journal of Chemical Education*, *92*(1), 58-65. doi: 10.1021/ed5002416

Muller, D. A., Bewes, J., Sharma, M. D., & Reimann, P. (2008). Saying the wrong thing: Improving learning with multimedia by including misconceptions. *Journal of Computer Assisted Learning*, *24*(2), 144-155. doi: 10.1111/j.1365-2729.2007.00248.x

Musacchio, G., Piangiamore, G. L., D'Addezio, G., Solarino, S., & Eva, E. (2015). "Scientist as a game": learning geoscience via competitive activities. *Annals of Geophysics*, *58*(3), S0328.

Olympiou, G., & Zacharia, Z. C. (2011). Blending physical and virtual manipulatives: An effort to improve students' conceptual understanding through science laboratory experimentation. *Science Education, 96*(1), 21-47. doi:10.1002/sce.20463

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., van Riesen, S. A., Kamp, E. T., … & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47-61. doi: 10.1016/j.edurev.2015.02.003

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*(2), 381-391.

Pine, K., Messer, D., & St. John, K. (2001). Children's misconceptions in primary science: a survey of teachers' views. *Research in Science & Technological Education*, *19*(1), 79-96. doi: 10.1080/02635140120046240

Pompea, S. M., Dokter, E. F., Walker, C. E., & Sparks, R. T. (2007, June). *Using misconceptions research in the design of optics instructional materials and teacher professional development programs*. Paper presented at ETOP conference, Ottawa, Canada.

Posner, G. J., Strike, K. A., Hewson, P. W. & Gertzog, W. A. (1982). Accomodation of a scientific conception: Toward a theory of conceptual change. *Science education, 66*(2), 211-227.

Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics*, *10*(11), 805-817.

Roehrig, G. & Garrow, S. (2007). The impact of teacher classroom practices on student achievement during the implementation of a reform-based chemistry curriculum, *International Journal of Science Education*, *29*(14), 1789-1811, doi:10.1080/09500690601091865

Rubenstein, D. J. (2000). Stimulating children's creativity and curiosity: Does content and medium matter?. *The Journal of Creative Behavior, 34*(1), 1-17.

Schmidt, H., Spaaij, G., & De Grave, W. (1988). Opsporen van misconcepties bij middelbare scholieren. *Tijdschrift voor Onderwijsresearch*, *13*(3), 129-140.

Sharp, J. G., & Kuerbis, P. (2006). Children's ideas about the solar system and the chaos in learning science. *Science Education*, *90*(1), 124-147.

Smiley, B. (2011). *Science education at Riverside middle school A case study* (Unpublished doctoral thesis). Arizona State University.

Smith, G. W., & Puntambekar, S. (2010). *Examining the combination of physical and virtual experiments in an inquiry science classroom.* Paper presented at the Conference on Computer Based Learning in Science, Warsaw, Poland.

So, W. M. W., Cheng, M. H. M., Kong, S. C., & Ching, N. Y. F. (2014). Views of primary science teachers towards the use of online resources to support the implementation of inquiry learning. *Education 3-13*, *42*(4), 386-401. doi: 10.1080/03004279.2012.710640

Toth, E. E., Morrow, B. L., & Ludvico, L. R. (2009). Designing blended inquiry learning in a laboratory context: A study of incorporating hands-on and virtual laboratories. *Innovative Higher Education*, *33*(5), 333-344. Doi: 10.1007/s10755-008-9087-7

Treagust, D. F. & Duit, R. (2008). Conceptual change: a discussion of theoretical, methodological and practical challenges for science education. *Cultural Studies of Science Education*, *3*, 297-328. doi: 10.1007/s11422-008-9090-4

Triona, L. M., & Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction*, *21*(2), 149-173.

Unal, S. (2008). Changing students' misconceptions of floating and sinking using hands-on activities. *Journal of Baltic Science Education*, *7*(3), 134-146.

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction, 11*, 381-419.

Waldrip, B., Prain, V., & Sellings, P. (2013). Explaining Newton's laws of motion: using student reasoning through representations to develop conceptual understanding. *Instructional Science*, *41*(1), 165-189. doi: 10.1007/s11251-012-9223-8

Weaver, G. C. (1998). Strategies in K-12 science instruction to promote conceptual change. *Science Education*, *82*(4), 455-472.

Winer, G. A., Rader, A. W., & Cottrell, J. E. (2003). Testing different interpretations for the mistaken belief that rays exit the eyes during vision. *The Journal of Psychology*, *137*(3), 243-261. doi: 10.1080/00223980309600612

Wysocki, B., McDonald, N., Fanto, M., & McEwen, T. (2013, March). *Designing STEM activities to complement neural development in children.* Paper presented at IEEE Integrated STEM Education Conference (ISEC), Princeton, New Jersey.

Yang, D. C., & Lin, Y. C. (2015). Assessing 10-to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test. *Educational Research*, *57*(4), 368-388. doi: 10.1080/00131881.2015.1085235

Yang, D., Streveler, R. A., Miller, R. L., Slotta, J. D., Matusovich, H. M., & Magana, A. J. (2012). Using computer-based online learning modules to promote conceptual change: Helping students understand difficult concepts in thermal and transport science. *International Journal of Engineering Education*, *28*(3), 686-700.

Zacharia, Z. C. (2007). Comparing and combining real and virtual experimentation: an effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning*, *23*(2), 120-132. doi: 10.1111/j.1365-2729.2006.00215.x

Zacharia, Z., & Anderson, O. R. (2003). The effects of an interactive computer-based simulation prior to performing a laboratory inquiry-based experiment on students' conceptual understanding of physics. *American Journal of Physics*, *71*(6), 618-629. doi: 10.1119/1.1566427

Zacharia, Z. C., Manoli, C., Xenofontos, N., De Jong, T., Pedaste, M., van Riesen, S. A., ... & Tsourlidaki, E. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: a literature review. *Educational Technology Research and Development*, *63*(2), 257-302. Doi: 10.1007/s11423-015-9370-0

Zacharia, Z. C., Loizou, E., & Papaevripidou, M. (2012). Is physicality an important aspect of learning through science experimentation among kindergarten students?. *Early Childhood Research Quarterly*, *27*(3), 447-457. doi: 10.1016/j.ecresq.2012.02.004

Zacharia, Z. C., & Olympiou, G. (2011). Physical versus virtual manipulative experimentation in physics learning. *Learning and Instruction*, *21*(3), 317-331. doi:10.1016/j.learninstruc.2010.03.001

Zacharia, Z. C., Olympiou, G., & Papaevripidou, M. (2008). Effects of experimenting with physical and virtual manipulatives on students' conceptual understanding in heat and temperature. *Journal of Research in Science Teaching*, *45*(9), 1021-1035. doi:10.1002/tea.20260

Zardetto-Smith, A. M., Mu, K., Phelps, C. L., Houtz, L. E., & Royeen, C. B. (2002). Brains rule! Fun = learning = neuroscience literacy. *The Neuroscientist*,*8*(5), 396-404. doi: 10.1177/107385802236965