# t·p·é*

## Une collection d'outils pour (s')initier à la publication en éducation et valoriser la transformation pédagogique

* transformer et publier en sciences de l'éducation

Ces outils sont organisés en trois volets :
→ accompagner / structurer pour les accompagnant·es
→ rédiger pour les équipes enseignantes débutant en recherche-action
→ s'enrichir / ressources une série de références et de guides

Ils ont été développés par la Chaire recherche-action sur l'innovation pédagogique de l'Université Paris Saclay et l'institut Villebon - *Georges Charpak*, en collaboration avec l'UQAM, et sont le fruit du travail de Marine Moyon, Frédéric Bouquet, Jeanne Parmentier, et Martin Riopel, et d'Emmanuel Ahr pour l'outil EVA.

L'exploration, la conception de la charte graphique et la mise en forme des outils ont été réalisées par Dalva Rospape et Marie Jouble.

Retrouvez tous les outils sur :
*https://cep.villebon-charpak.fr/tpe*

Institut Villebon Georges Charpak · université PARIS-SACLAY · UQÀM

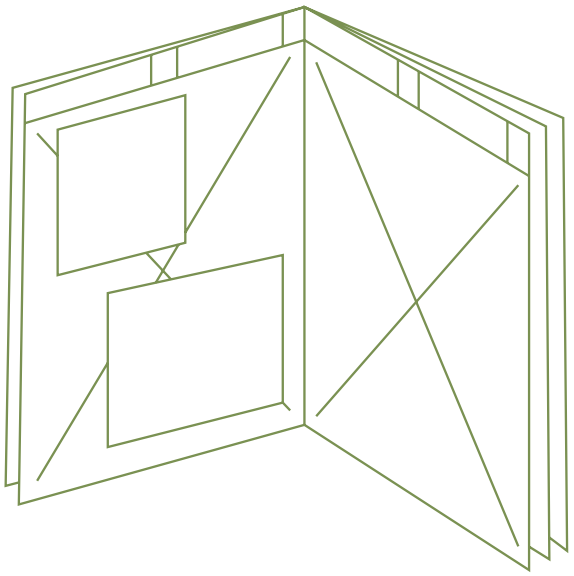
# → rédiger

## article en kit

Cet outil s'adresse aux personnes enseignantes porteuses d'un projet de recherche-action.

Il vise à les accompagner dans les premières étapes de la rédaction de leur article scientifique.

L'outil se présente sous la forme de vignettes, extraites d'articles scientifiques. Chaque vignette propose un exemple de formulation, pouvant être adapté au contexte spécifique du projet. Il offre ainsi un appui concret pour amorcer l'écriture. Ces exemples sont à utiliser comme des modèles d'inspiration, dans le respect des règles de l'intégrité académique, sans reproduction littérale.



DOI : 10.1002/articlefakeSoTL | Received / / | Revised / / | Published / / | JKIT / WILEY

research article

# titre de mon article factice

auteur·ices de l'article

université

correspondance

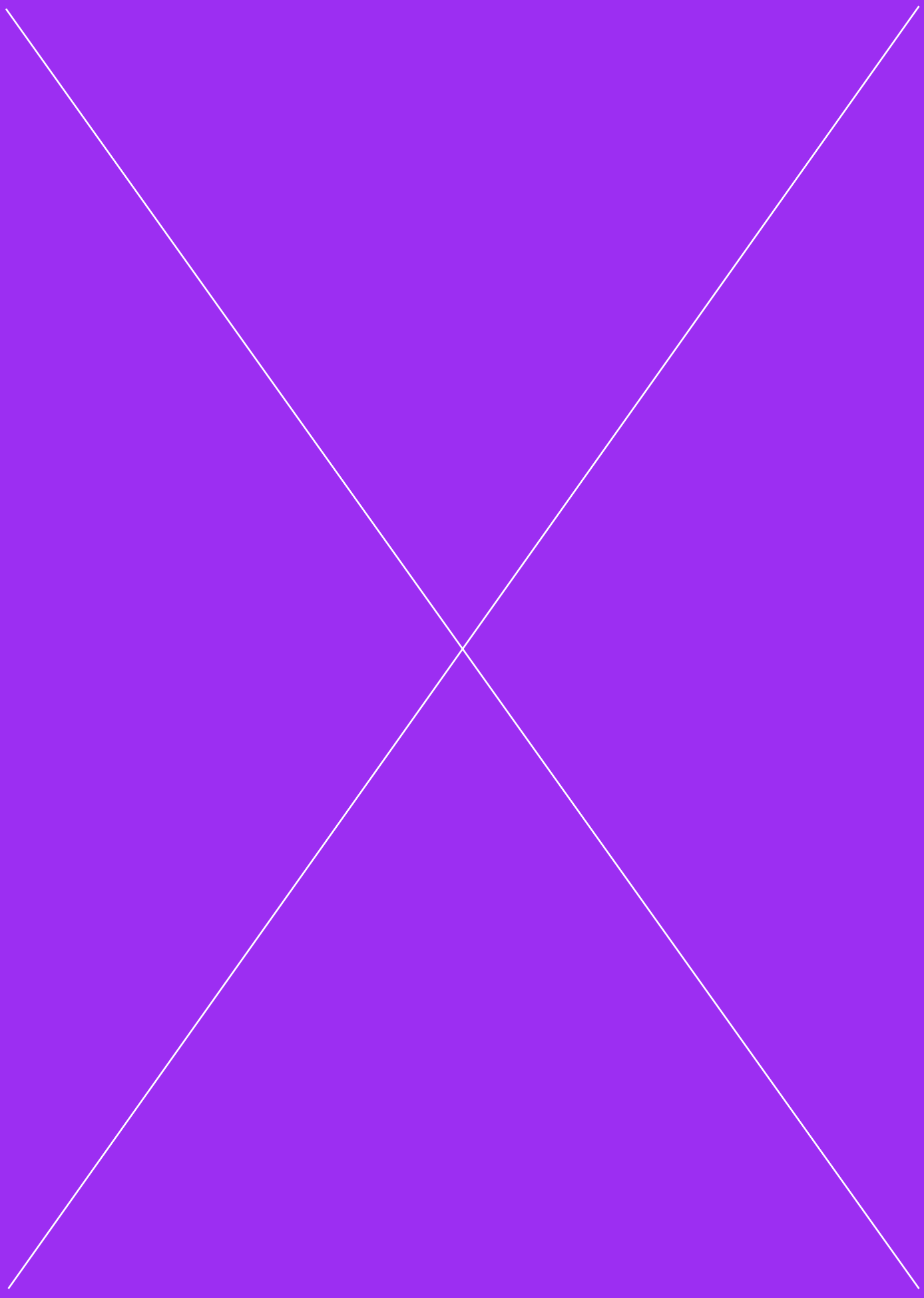
## résumé

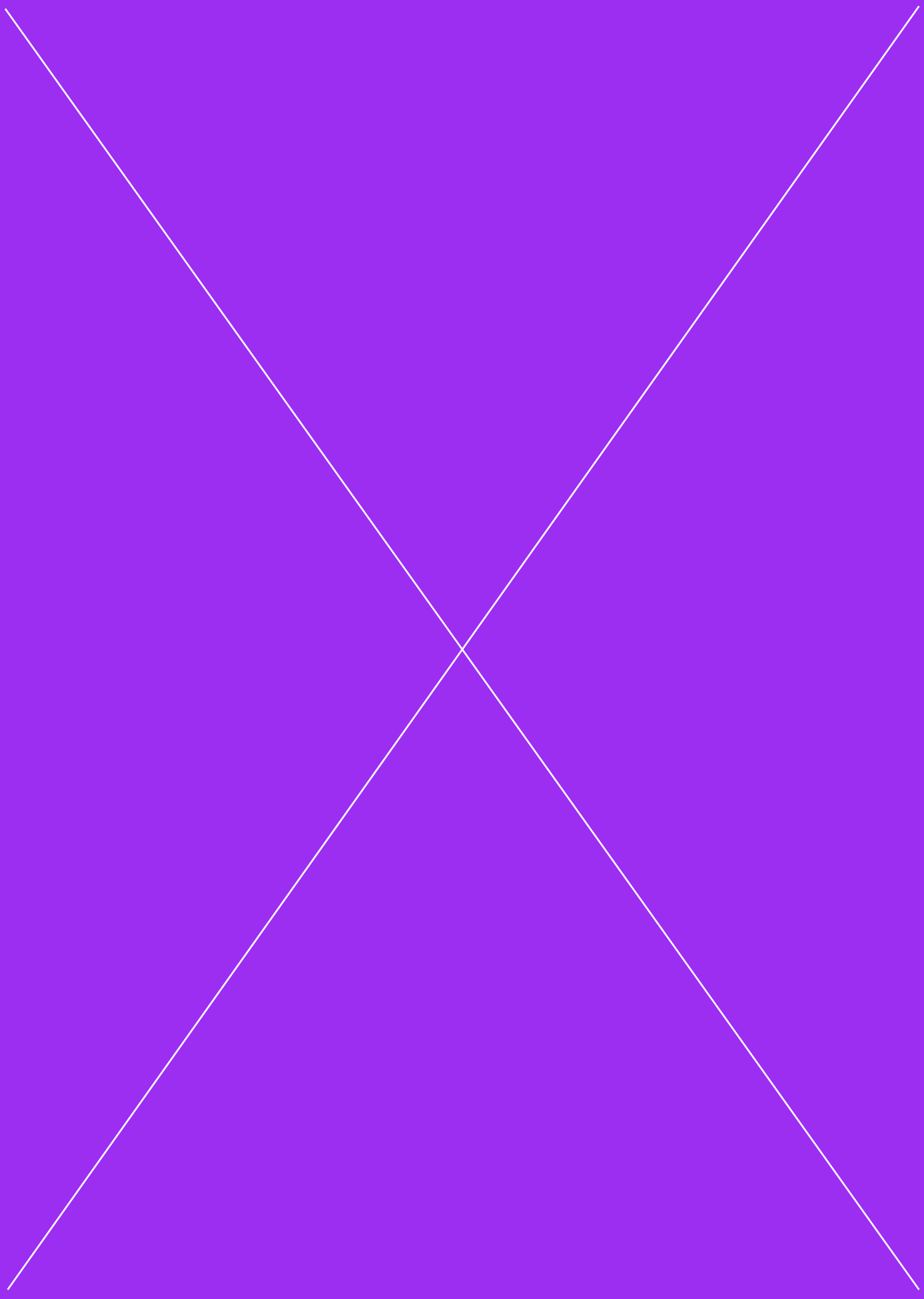
contexte

manifestation du problème

emergence d'un besoin

| ↓ | résumé |
| --- | --- |
| pertinence scientifique | |
| objectif de l'étude | |
| méthodologie | |
| résultats principaux | |
| discussion / conclusion | |





# introduction → objectifs / questions de recherche / hypothèses

méthodologie → participant·es

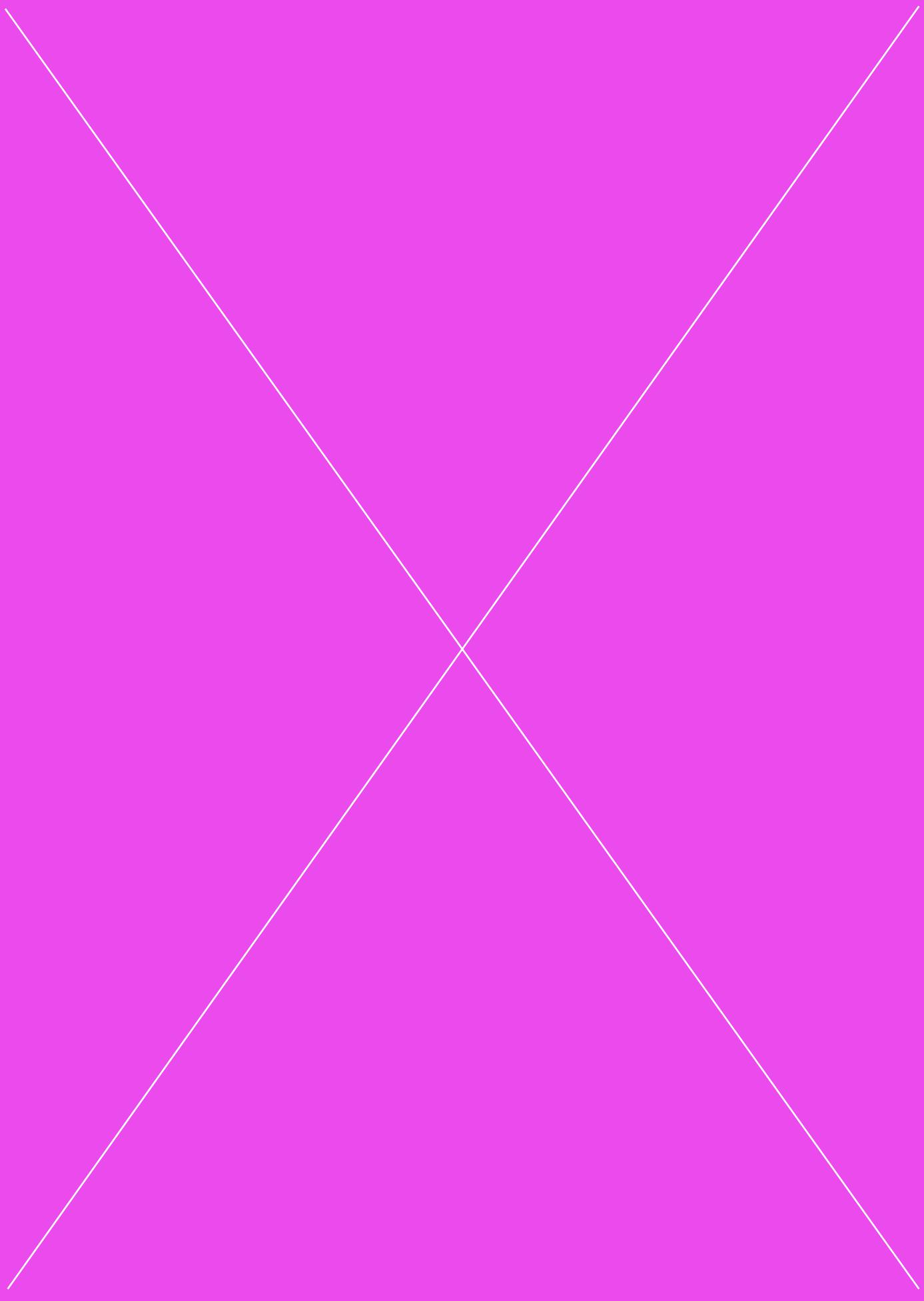
méthodologie → protocole

méthodologie → dispositif

méthodologie → analyse de données

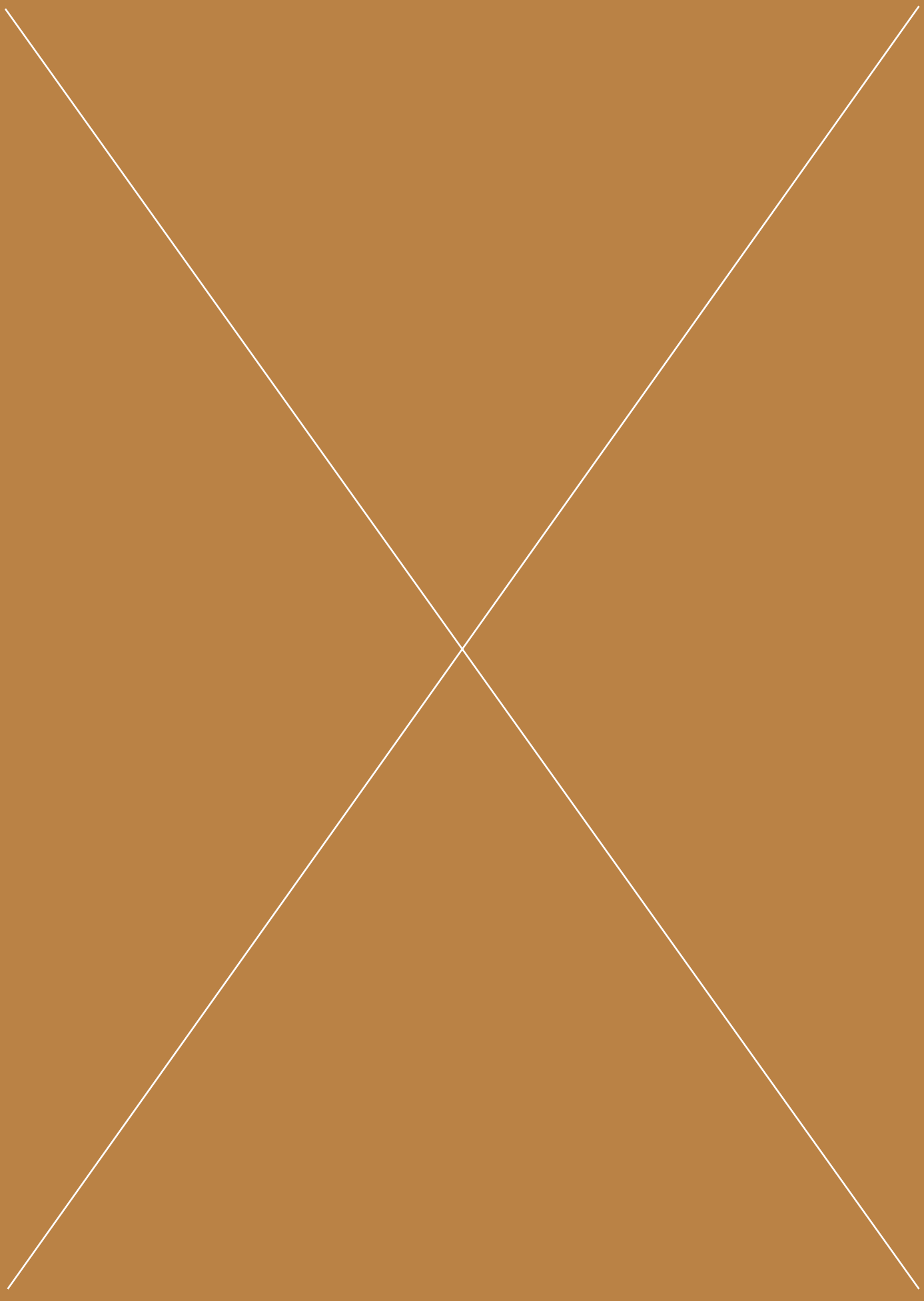
méthodologie → éthique

# résultats



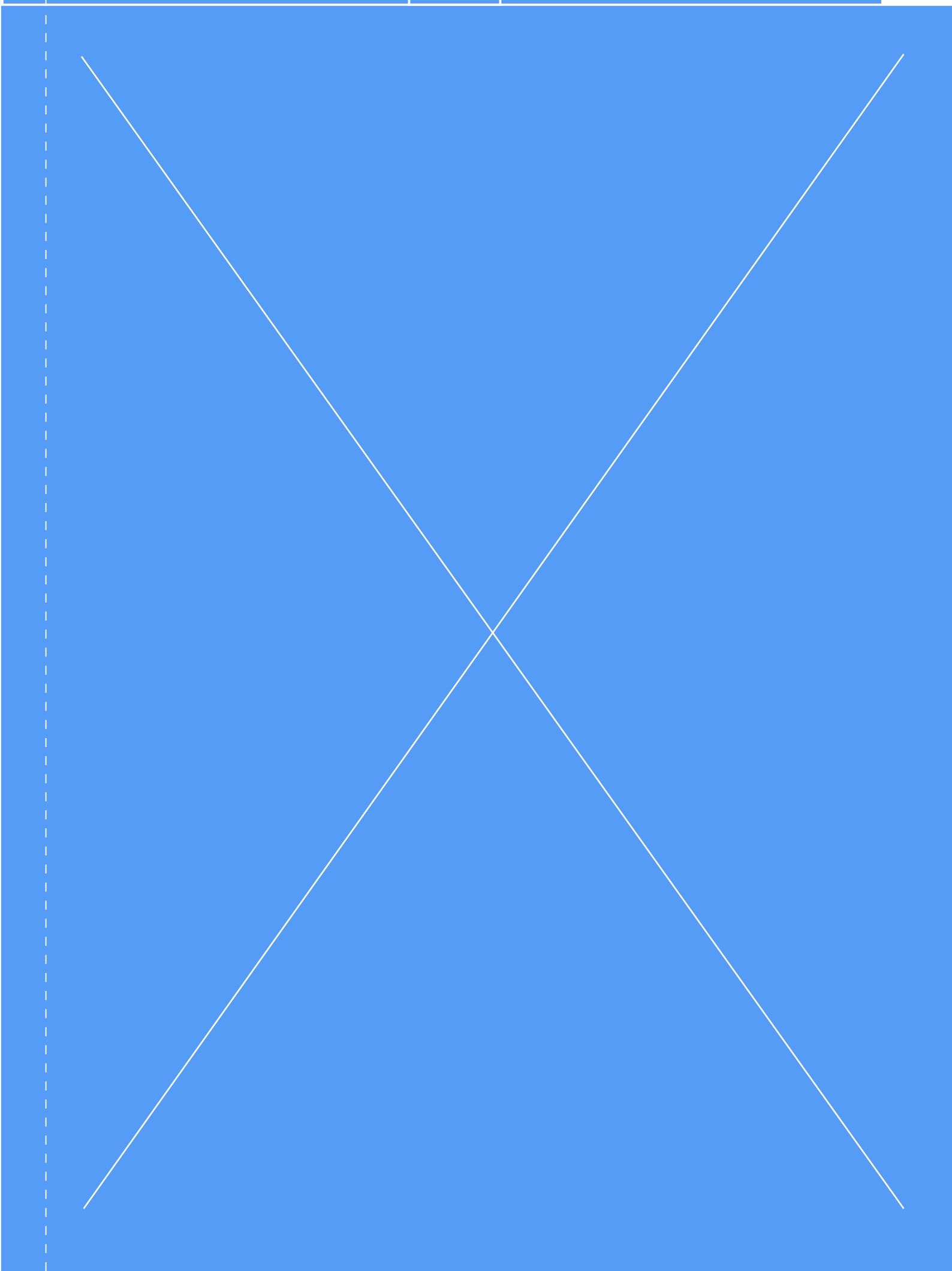
# discussion

# conclusion



# références bibliographiques

# annexes



Institut Villebon Georges Charpak

université PARIS-SACLAY

UQÀM

# résumé


With developments in technology (e.g., "Web 2.0" sites that allow users to author and create media content) and the removal of publication barriers, the quality of science information online now varies vastly. These changes in the review of published science information, along with increased facility of information distribution, have resulted in the spread of misinformation about science. As such, the role of evaluation when reading scientific claims has become a pressing issue when educating students. While recent studies have examined educational strategies for supporting evaluation of sources and plausibility of claims, there is little extant work on supporting students in critiquing the claims for flawed scientific reasoning. This study tested the efficacy of a structured reading support intervention for evaluation and critique on cultivating a critical awareness of flawed scientific claims in an online setting. We developed and validated a questionnaire to measure epistemic vigilance, implemented a large-scale (N = 1081) Randomized Controlled Trial (RCT) of an original reading activity that elicits evaluation and critique of scientific claims, and measured whether the intervention increased epistemic vigilance of misinformation. Our RCT results suggested a moderate effect in students who complied with the treatment intervention. Furthermore, analyses of heterogeneous effects suggested that the intervention effects were driven by 11th-grade students and students who self-reported a moderate trust in science and medicine. Our findings point to the need for additional opportunities and instruction for students on critiquing scientific claims and the nature of specific errors in scientific reasoning.

(Tseng et al., 2021)


[contexte]

[manifestation du problème]

[émergence d'un besoin]

[pertinence scientifique]

[objectif de l'étude]

[méthodologie]

[résultats principaux]

[discussion / conclusion]

# résumé

[contexte]

[émergence d'un besoin]

[pertinence scientifique]

[méthodologie]

[résultats principaux]

[discussion / conclusion]

Serious games have become increasingly available to educators. Empirical studies and meta-analyses have examined their impact on learning achievement. However, natural sciences could have a special relation to serious games by their systematic use of quantitative and predictive models that can generate microworlds and simulations. Since no known meta-analysis on serious games observed a significant impact in the specific context of science learning, the present meta-analysis synthesised results from 79 empirical studies that compared the impact on science learning achievement of instruction using serious games versus instruction using more conventional methods. Consistent with theory and past meta-analyses not specifically related to science learning, post- instruction learning achievement was weakly to moderately higher for declarative knowledge, knowledge retention and procedural knowledge for students taught with serious games. Furthermore, findings of the present work suggest that five moderator variables produced significant effects on the relationship between playing serious games and learning outcomes, and three showed consis- tent variations in mean effect size that could lead to significance, with more studies and larger samples. These findings are discussed in connection with previous meta-analyses' findings, potential ped- agogical implications and possible future research.

(Riopel et al., 2019)

# résumé

The purpose of the present study is to examine the effects of IMPROVE, a meta-cognitive instructional method, on students' mathematical knowledge, mathematical reasoning and meta-cognition. [objectif de l'étude]

Participants were 81 students who studied a pre-college course in mathematics. Students were randomly assigned into one of two groups and groups were randomly assigned into one of two conditions: IMPROVE vs. traditional instruction (the control group). Both groups were exposed to the same learning materials, solved exactly the same mathematical problems, and were taught by the same experienced teacher. The IMPROVE students were explicitly trained to activate meta-cognitive processes during the solution of mathematical problems. The control group was exposed to traditional instruction with no explicit exposure to meta-cognitive training. [méthodologie]

Results indicate that the IMPROVE students significantly outperformed their counterparts on both mathematical knowledge and mathematical reasoning. In addition, the IMPROVE students attained significantly higher scores then the control group on the three measures of meta-cognition: (a) general knowledge of cognition; (b) regulation of general cognition; and (c) domain-specific meta-cognitive knowledge. [résultats principaux]

The theoretical and practical implications are discussed. [discussion / conclusion]

(Meravech et Fridkin, 2006)

[contexte] Many noncognitive constructs affect mathematical problem-solving performance.

[objectif de l'étude] The aim of the present study is to investigate the direct and indirect effects of a number noncognitive constructs such as mathematics self-efficacy, mathematics anxiety, and metacognitive experience on the mathematical problem solving of middle-school students.

[méthodologie] The sample consisted of 517 seventh-grade Turkish students of whom 252 were male (49%) and 265 were females (51%). The instruments used in this study were a mathematical problem-solving performance test, a mathematics self-efficacy scale, a mathematics anxiety scale, a metacognitive experience scale, and a mathematics motivation scale. Two-stage structural equation modeling was used to exam- ine the relationships between the noncognitive contructs and problem solving.

[résultats principaux] Metacognitive experience was the only noncognitive construct, which had a direct effect on mathematical problem-solving performance; it also mediated the effects of self-efficacy, motivation, and math- ematics anxiety on performance. Motivation and mathematics anxiety had an indirect effect on mathematical problem-solving performance through self-efficacy.

(Özcan et Eren Gümüs, 2019)

Important issues are associated to science and technology (S&T) education in modern societies. First among these is the development of a scientific and technological culture for all citizens, whether they are destined to pursue careers in S&T or not. In societies strongly marked by scientific knowledge and technological advances, the paucity of such a culture in the population hinders the exercise of informed citizenship. These issues are also related to social progress. The shortage of people with training in this field may deprive societies of critical human resources needed for the industrial and economic development on which they are based.

(Hasni et Potvin, 2015)

Traditionally, there has been a gap between what students have learned and the skills that they have acquired in the university and what companies have required when hiring new employees (Moore & Morton, 2017; Pang, Wong, Leung, & Coombes, 2019; Hayter & Parker, 2019). Despite having identified this gap, many universities continue using traditional learning methodologies focused on the lecturer rather than on the student, thus hindering the development of essential skills required in the workplace (Chaudhry & Rasool, 2012; Lai, Hsiao, & Hsieh, 2018; Pelger & Nilsson, 2018).

In contrast to this generalized stream of teaching practice in higher education, the first benefit of this research is the providing of guidelines for implementing a successful active learning setup in the university context, centered on the student and particularly effective in compensating for the difference between knowledge and skills that characterizes most of the teaching and learning methodologies still commonly used in higher education.

The Assessment and Teaching of 21st Century Skills (ATC21S; Care, Griffin, & Wilson, 2018), the Bologna process and the European Higher Education Area (EHEA; Zahavi & Friedman, 2019), or the Partnership for 21st Century Learning (P21; Van Laar, van Deursen, Van Dijk, & de Haan, 2017) form part of an international movement focusing on conceptual learning frameworks, oriented toward the skills required for students to succeed in a fast-changing digital society. In this context, the engagement of the student with the teaching-learning process plays a fundamental role (Boekaerts, 2016; Guo, 2018; Lei, Clemente, & Hu, 2019). Engagement is helped by any of these 21st century skills frameworks awarding students with an active role in their own learning. In doing so, they can acquire a series of abilities also associated with content-knowledge learning that will make them more employable when leaving the university (Daellenbach, 2018; Fletcher, Sharif, & Haw, 2017).

Following this international movement, in recent years, there has been a change in the way lectures in higher education are being delivered, going from the traditional instructor-based teaching model to active and student-centered learning experiences that generate engagement and contribute both to the acquisition of knowledge and the skills necessary to enter into the labor market. Game-based learning (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Sousa & Rocha, 2019), cooperative learning (Azizan, Mellon, Ramli, & Yusup, 2018; Johnson & Johnson, 2009), problem-based learning (Loyens, Jones, Mikkers, & van Gog, 2015; Schmidt, Van der Molen, Te Winkel, & Wijnen, 2009), and the flipped classroom (Awidi & Paynter, 2019; Lage, Platt, & Treglia, 2000) are some of the most remarkable examples.

(Murillo-Zamorano et al., 2019)

Programming and computational thinking (Wing, 2006) constitute a new literacy for all citizens in a digital society. Programming involves skills such as logical and systematic thinking, which can be used to solve problems encountered in various learning contexts or in daily life, beyond the professional field of computer science (Tsai et al., 2019). For this reason, in recent years, programming skills have been included in the school curriculum standards of several countries (e.g. in Canada: Government of Quebec [Gouvernement du Quebec], 2018; in Taiwan: Ministry of Education, 2016; in the United States: National Science Foundation [NSF], 2016). One challenge faced in teaching programming in K-12 settings is to address the persistent underrepresentation of women in computer science, which is an important topic for economic and social justice reasons (Beyer, 2014). Computer science has one of the lowest shares of women degree recipients among science, technology, engineering and mathematics (STEM) fields, and the share of women receiving bachelor's and doctorate degrees has declined over time (NSF, 2019). In the USA, only 19% of computer science degrees were awarded to women in 2016 compared to 27% in 1997 (NSF, 2019), and only 0.4% of female first-year university students intended to major in computer science compared to 2.9% of their male counterparts (NSF, 2013). In Canada, the trend is comparable, in that women account for only 30% of students who graduated from mathematics and computer science programmes in the last decade (Hango, 2013) with Québec ranging below average at 19% of female computer science undergraduates (Chaire pour les femmes en sciences et en genie au Québec, 2021).

(Allaire-Duquette et al., 2022)

In recent decades, there has been a shift in education from a transmissive to a constructivist model that emphasises learners' activity, including their own practical work. In this paper, we follow the definition given by Millar (2010), who refers to practical work as 'any science teaching and learning activity in which the students, working individually or in small groups, observe and/or manipulate the objects or materials they are studying'. Traditionally, practical work has been expected to have many functions – among others, enhancing conceptual learning, stimulating interest, understanding the nature of science or gaining insight into scientific methods and approaches. Further we limit ourselves to the affective aspects of practical work.

Practical work is often used by teachers to present science 'in a better light', as an engaging and enjoyable subject (Abrahams, 2007). Students really prefer practical work to other forms of instruction (Bennett, 2003; Owen et al., 2008), though the reasons for this are not crystal clear (Bennett, 2003). Some authors point out that practical work allows students to interact with their teacher and classmates in less serious atmosphere and to manage their work at their own pace (Gardner & Gauld, 1990). According to Bennett (2003), practical work may serve as an escape from the boring routine of writing, reading, and listening to the teacher. Anyway, for students to enjoy and value practical work, its purpose must be clear (Bennett, 2003) and its assignment should provide them with an appropriate challenge as well as some control over what they have to do (Bennett, 2003; Hodson, 1990).

(Kácovský et al., 2023)

With the advent of recent Web technologies, the landscape of information available to the public has changed dramatically. While more information is available than ever, there is a greater amount of false or misleading information that is published, distributed and consumed by the public (Barzilai & Chinn, 2020; Bromme & Goldman, 2014; Graesser, Millis, D'Mello, & Hu, 2014; Kiili, Laurinen, & Marttunen, 2008; Lewandowsky, 2020; Rojecki & Meraz, 2016). New web technologies have facilitated these issues, allowing scientific misinformation to spread widely and rapidly without the gatekeeping of traditional media publication standards (Moran, 2020; Zarocostas, 2020).

(Tseng et al., 2021)

Over the past decade, the number of sustainability programs in higher education has grown significantly. In the USA, for example, between 2012 and 2016, this number increased by 15% to 2361, offered by 872 institutions (Vincent et al. 2017). These include sustainability-focused programs, reflecting the field of sustainability science (Kates 2011; Yarime et al. 2012; Lang et al. 2012; Wiek et al. 2015) and sustainability-oriented programs in business, education, law, and so forth. Graduates from these programs seek employment as sustainability professionals or sustainability researchers. Projections for the USA suggest up to 9% growth in the existing sustainability labor market through 2024 (Johnson et al. 2019).

(Brundiers et al., 2021)

Changes in science and technology have affected the structure of societies and have led to rapid change in human profile. In order to adapt to the changing human profile, reforms in education as well as scientific and technological enrichment in educational environments have become necessary. In an era where the use of technology in daily life is rapidly increasing, effective use of technology in education has become important (Alkan, 1998). Since abstract concepts are frequently used in explaining nature and natural phenomena during science courses, supporting these courses with technology has become a necessity. In order to enrich students' learning environments, it is important to amplify their visual and intellectual engagement through the use of technology, especially when explaining abstract and difficult concepts. The use of technology also allows students to perceive phenomena in science courses in a multidimensional manner, to interpret information better and to keep their attention on the course (Akpinar, Aktamis, & Ergin, 2005). Educational technologies in science teaching serve to improve the quality of science courses through effective scientific activities, develop students' reasoning skills in science courses, help students discover knowledge, enhance their problem-solving abilities and convey difficult-to-comprehend situations that are difficult to recognize in daily life (Karamustafaoglu, Cakir, & Topuz, 2012). It is important to provide technology-based instructional materials for students in science courses and to reorganize learning environments in accordance with students' needs, thereby enabling them to learn through action and experience (Akpinar, Aktamis, & Ergin, 2005). An emergent technology, augmented reality (AR) was used in the study. Thanks to AR technology, it has become possible to prepare effective and interesting technology-based instructional materials.

(Sahin et Yilmaz, 2020)

Serious games for science learning

It has been widely argued that more active instructional methods must be implemented in science classrooms (Avvisati, 2011; Aziz, Nor, & Rahmat, 2011; Millar, 2011; Wieman, 2012). To define these more active instructional methods, one can refer to the general proposition of Bonwell (1991, p. iii) for students to be engaged in usual tasks such as listening, reading, writing, discussing or solving problems but also, and most importantly, to be engaged in higher-order thinking tasks such as analysis, synthesis, or evaluation and for which an inquiry-based approach to science produces good examples. The National Research Council (National Research Council, 2011, p. 22) has also pointed out that: A growing body of research indicates that engaging students in science processes (inquiry) can motivate and support science learning. However, because inquiry approaches can be difficult for students, teachers, and schools, they are rarely implemented.

In this particular context, computer simulation and serious games can play a special role because the quantitative and predictive models of science can be used to generate interactive microworlds and simulations that can be freely experienced. 'Computer simulations and games have great potential to catalyze and support inquiry-based approaches to science instruction, overcoming curricular and logistical barriers.' (National Research Council, 2011, p. 22).

(Riopel et al., 2019)

Students' utilization of mathematical knowledge in their daily lives, ability to think mathematically, comprehension of problem-solving strategies as well as utilization of these strategies are emphasized in curricula in many countries (Australian Education Council, 1991; National Council of Teacher of Mathematics, 2000; Turkey's Ministry of National Education (MONE), 2016; Victorian Curriculum Assessment Authority, 2016). Despite this emphasis on mathematical problem solving in learning and instruction curricula, student performance in this area is declining in a number of countries (Organization for Economic Co-operation and Development (OECD), 2016). International assessments place Turkish students at the end of compulsory education (age 15–16) below the average for students across OECD countries (OECD, 2016). Furthermore, Turkish students' mathematics performance in the Program of International Student Assessment (PISA) 2015 was inferior to that in PISA 2009 and 2012 (Taş, Arıcı, Ozarkan, & Özgürlük, 2016). Concerns about mathematics performance are not restricted to developing countries, however; mathematics performance is widely considered a major problem even in developed countries such as Australia and the United States of America (Lamb & Fullarton, 2002; OECD, 2016; Thomson, De Bortoli, & Underwood, 2017).

(Özcan et Eren Gümüş, 2019)

Introduction

Physics classes often involve problems that are technical or presented in purely scientific contexts (Häussler, Hoffmann, Langeheine, Rost, & Sievers, 1998; Holmes, Burns, Marra, Stubbe, & Vine, 2003, Murphy, 1990; Srivastava, 1996) which may explain the gender disparity in the field and women's lack of interest in the subject (Murphy & Whitelegg, 2006). The underrepresentation of women in physics at university (see Figure 1) has been widely addressed in American and international literature (eg. Baram-Tsabari & Yarden, 2008 ; Lavonen, Byman, Juuti, Meisalo, & Uitto, 2005). Although the gender gap in other scientific fields has been bridged in recent years, the gender disparity in physics, in both professional and academic spheres, is yet to be reduced (Ivie & Stowe, 2000 ; Lorenzo, Crouch, & Mazur, 2006). Among all sociological, psychological, and cultural factors affecting the choice of women to pursue studies and careers in physics, there is a consensus that one of the major contributing factors is that women develop significantly lower levels of interest in physics than men (see, for example, related studies in Scotland (Stark & Gray, 1999), Australia (Dawson, 2000), the United States (Jones, Howe, & Rua, 2000), England (Murphy and Whitelegg, 2006), Germany (Hoffmann, 2002), and an international study (Lavonen et al., 2005)).

Although the desire to engage in learning or problem solving is essential in an individual's learning process (Hidi & Harackiewicz, 2000), teachers rarely have a clear understanding of their roles in developing their students' interest toward their discipline (Lipstein & Renninger, 2006), and this seems to be the case with physics.

Hoffmann (2002) found that less than 20% of the variance in the interest of men and women in physics was due to content, whereas the remaining 80% was related to the context in which the concepts and activities were presented to the students. For example, classes in mechanics rarely focus on the study of phenomena in context (Murphy and Whitelegg, 2006), but use sports, military, and automotive contexts (Hoffmann, 2002). To stimulate and maintain the interest of an individual in a particular subject matter, a teacher needs to initiate a first critical phase.

(Hidi & Renninger, 2006) that is accompanied by a sudden change in the emotional and cognitive processes (Hidi & Baird, 1986, 1988, Mitchell, 1993). Stimuli-engaging emotions are then assumed to develop students' interest in physics. Therefore, physics teachers must focus on stimulating the emotions of their students. Certain contexts are suggested to be more effective in stimulating emotions (Duit, Häussler, Lauterbach, Mikelskis, & Walter, 1992, p.109) : [...]

Therefore, strategies for stimulating students' interest and engaging them in physics must be identified, considering that physics, as it is traditionally taught, is perceived by students as irrelevant and demotivating (Murphy & Whitelegg, 2006). In fact,a promising educational tool could be developed if science teachers were to create and maintain girls' situational interest in science through the introduction of non-scientific contexts and illustrations that are known to activate girls' individual interests (Krapp, 1998; Mitchell, 1993). (Kerger, Martin, & Bruner, 2011, p.608-609)

(Allaire-Duquette et al., 2014)

Enhancing student interest in science and increasing student self-efficacy in order to increase expectancy and improve flagging attitudes towards science are all foundational in building the U.S. STEM economy (Osborne et al., 2003). However, interest in science generally declines as students progress from elementary to secondary school (Höft et al., 2018; Potvin & Hasni, 2014), and perceptions of gender and racial/ethnic barriers to STEM are prevalent (Grossman & Porche, 2014), further limiting the future STEM talent pool. As interest and self-efficacy are both predictors of future choices and career paths (Chemers et al., 2011; Kelly et al., 2013), creative, transferable approaches are needed to promote and sustain student interest and self-efficacy and improve students' perceptions of who can be a scientist.

(Howell et al., 2023)

Students at primary school have difficulty fully comprehending complex abstract concepts. For example, basic astronomy concepts are abstract in nature, which interferes with students' comprehension of the material and negatively affects their attitudes towards the course (Gundogdu, 2014). In order to overcome these difficulties, it is necessary to make abstract concepts in science more concrete through the use of visuals in teaching. As a result, a more meaningful learning environment can be designed in which enhanced achievement and the development of positive attitudes can be expected. AR plays an important role in embodying and visualizing abstract concepts in accordance with students' comprehension levels, and in enabling the observation of phenomena that are impossible to encounter in real life (Arici, Yildirim, Caliklar, & Yilmaz, 2019). Considering the importance of AR applications, this study aims to investigate whether AR technology affects middle school students' achievement and attitudes during their science courses.

(Sahin et Yilmaz, 2020)

Furthermore, over the past decades a growing gap has been observed between the scientific and technical expertise offered by schools, on one hand, and the social demand in this regard, on the other: societies are showing a growing need for individuals trained in this field, while the number of students attracted to it is stagnating and in some cases declining (Organisation for Economic Cooperation and Development [OECD], 2006; 2008). This gap, which many describe as students' loss of interest in S&T, has been observed in many parts of the world, for example in England (Convert, 2005; Cotgreave & Davies, 2005; Hannover & Kessels, 2004) in Germany (Haas, 2005), in the United States (Foster, 2010), in Canada (Dobson & Burke, 2013) and also in France (Convert, 2005; Ourisson, 2002 ; Porchet, 2002).

In this context, developing students' interest in S&T and its related studies and careers, over and beyond the quality of learning, must be a preoccupation for schools, educational policies and academic research in education.

(Hasni et Potvin, 2015)

However, it remains a challenge for employers, students, educators, and program administrators to clearly articulate what competencies these programs develop in students (Barth et al. 2007; Rieckmann 2012). Several frameworks of competencies in sustainability have been proposed, with the most commonly referenced one being the framework of key competencies in sustainability by Wiek et al. (2011). Despite this emerging convergence, a state of clarity has yet to be achieved. A first challenge is the variety of terms still in use for similar competencies, creating a "sea of labels" and resulting in "terminological confusion" (Sterling et al. 2017, p. 153, Shephard et al. 2018). A second challenge is that new proposals for sustainability competencies continue to be presented as lists of items (Wilhelm et al. 2019), although scholars acknowledge the importance of a framework as a set of distinct, yet interrelated competencies (Wals 2015; Glasser and Hirsh 2016; Engle et al. 2017). Lastly, there is no explicit consensus on a specific framework of key competencies in sustainability.

This lack of clarity has several negative effects. Sustainability programs often do not clearly articulate the learning objectives for their students (O'Byrne et al. 2015). Prospective students struggle to compare sustainability programs as they decide to which program to apply. Instructors lack guidance on what competencies to convey to students. Graduates of sustainability programs encounter difficulties in articulating their competencies while employers lack a trustworthy reference to compare candidates' profiles (Barber 2016). In the absence of commonly agreed upon key competencies in sustainability and related program-level learning objectives, accrediting bodies are unable to assess learning and benchmark degree programs (Vare et al. 2019), making systematic comparison and evaluation of degree programs difficult. This absence is at odds with the fact that sustainability is recognized as an established academic field (Kates 2011; Lang et al. 2012; Yarime et al. 2012; Miller et al. 2014; Wiek et al. 2015) and sustainability practice as a profession (ISSP 2020), respectively.

Considering these challenges, it is time to work toward "broadly acceptable, detailed descriptions" of key competencies in sustainability to provide "guidance for program and curriculum development or major re-organization of academic institutions" (Glasser and Hirsh 2016, p. 132). Creating a shared frame of reference and quality standards enables credibility and professional trust in sustainability programs. It would provide a shared language around program-level learning objectives, facilitating comparability of programs, and an increased understanding established academic field (Kates 2011; Lang et al. 2012;

(Brundiers et al., 2021)

Since the 2000s, studies have reported issues involving elementary science education and the refractory or negative attitude of teachers toward teaching science and technology (S&T) (Cavallo et al., 2002; Cavas et al., 2013; Christidou, 2011; Denessen et al., 2015; Van Aalderen-smeets et al., 2012). Research reveals that many teachers demonstrate limited scientific knowledge (Epstein & Miller, 2011; Palmer, 2004; Van Driel et al., 2014) and feel unconvinced about their ability to teach S&T (Epstein & Miller, 2011; King et al., 2001). Some researchers also report that teachers lack instructional methods, preparation time, and material for conducting experiments (Carleton et al., 2008; Lumpe et al., 2000). Teachers then adopt a negative attitude toward teaching S&T, sometimes avoiding it in favor of teaching more basic subjects (Hasni, 2005; Kazempour, 2014), or not teaching it at all (Appleton & Kindt, 1999; Chen et al., 2014; Goodrum et al., 2001; Palmer, 2004).

(Marec et al., 2021)

In the literature, there are many studies on the use of AR in science education (Arici, Yildirim, Caliklar, & Yilmaz, 2019). For example, Shelton and Hedley (2002) observed AR's effect on undergraduate Geography students' comprehension of Earth-Sun relationships. At the end of the study, it was concluded that AR technology positively affected students' understanding of the subject. Another study, carried out by Kerawalla, Luckin, Seljeflot, and Woolard (2006), investigated the impact of AR on dialogues between teachers and students. They found that the experimental group which learnt via AR had better dialogues with their teacher and had to warned less about poor behavior during their lessons as they were more focused than usual. Additionally, Wang and Chi (2012) investigated the effects of AR on students' satisfaction and achievement. Their data demonstrated that students' achievement had improved and that they were satisfied with the AR system. Abdüsselam (2014) stated that AR technology would be useful in teaching magnetism in Physics courses. It was revealed that AR technology could enable visualization of the magnetic field and contribute to the concretization of the subject. Additionally, a study conducted by Wojciechowski and Cellary (2013) showed that students found AR-aided classes entertaining and enjoyed the application. They also found a strong positive relationship between attitudes and the use of AR applications.

Di Serio, Ibàñez, and Delgado-Kloos (2013) investigated the effects of AR technology on students' motivation. Results indicated that AR had a positive effect on middle school students' motivation. Delello (2014) studied prospective teachers' perception on the use of AR in science courses. In this study, it was found that AR applications positively affected the learning environment by increasing motivation, class commitment and the teacher's excitement, and enabled the partnership in the implementation of the application. Cai, Wang and Chiang's (2014) study also concluded that AR, as a computer-aided tool, had considerable integrative effects on learning. In another study, Erbas, 2016 used a mobile AR application on tablet computers in a 9th grade biology class. He investigated these students' academic achievements and attitudes towards the course. It was concluded that the use of AR technology in courses enhanced student achievement and improved their attitudes towards the course. Yildirim, 2016 investigated the effect of AR applications on students' achievement, motivation, perception of their own problem-solving skills and attitudes. In this study, there was two experimental groups which are students who learnt via computer-based AR applications and students who learnt via computer-based AR applications. Students who learnt via computer-based AR applications were more successful than those who learnt via tablet-based applications. The study also demonstrated that both experimental groups had higher motivation levels than the control group. Summarizing the studies about AR in the literature, it is seen that students' attitudes, motivations, interests and achievements are frequently discussed. In general, a single variable was considered and there were few studies dealing with multiple variables. Additionally, although various scientific topics have been covered in investigations on the impact of AR, the solar system has not yet been investigated in detail. Accordingly, this study aims to fill this gap in the literature.

(Sahin et Yilmaz, 2020)

The increasing focus on argumentation as a topic in science education research indicates its growing importance among scholars (see, for example, Asterhan & Schwarz, 2007; Kuhn & Crowell, 2011; Özdem et al., 2017 and references therein). By its analytical, dialectical, and rhetorical nature, argumentation is an essential skill in learning to solve various kinds of problems (Jonassen, 2011). One aspect which illustrates the significant role of argumentation skills in science education stems from the fact that science generally advances by argumentation, dialogue, and revolutionary ideas than by doctrinaire (Popper, 1965; Voss, 2006). In this instance, Von Aufschnaiter and colleagues (2008), suggested three reasons why students, in particular science students, should be exposed to argumentation: (a) scientists engage in argumentation to develop and improve scientific knowledge, (b) the public has to use argumentation to engage in scientific debates, and (c) students' learning of science requires argumentation (p. 102). For this reason, it is important that students of all ages engage in argumentation and develop their argumentation and reasoning skills to gain a better understanding of science, themselves, others, and the world (Özdem et al., 2017). Although the importance of argumentation in the development of scientific knowledge and solving of problems encountered in everyday life has long been recognized by researchers (Albe, 2008; Belland et al., 2011; Jonassen, 2007; Spector & Park, 2012), more research is needed into the framework of dialogical argumentation-based instruction (DAI) which provides a very different perspective on students who are in a dialogical relation with others, and who contribute to a conversation by means of thinking, sense making, reasoning, and problem solving in the science classroom. One line of argument deals with the need to engage students in activities in which argumentation structure depicts how reasoning is used in relation to solving ill-defined science problems. Unlike well-defined problems (WDPs) which can be solved with a high degree of certainty and the solution is agreed upon by experts, illdefined problems (IDPs) often possess multiple solutions or unclear answers, and thus the student has to examine different possibilities, assumptions, and evaluate possible solution outcomes (Iwuanyanwu, 2020; Jonassen, 2011). At the end, when a solution is proposed, it usually is justified by arguments and/or counterarguments that indicate why the solution is reasonable (Voss, 2006). Thus, the interplay between students developing the ability to solve IDPs (Shekoyan & Etkina, 2007) and acquiring the concepts of science through a dialogical argumentation-based instruction- DAI (Asterhan & Schwartz, 2007; Iwuanyanwu & Ogunniyi, 2020) – one building upon the other – is indispensable in successful science learning. Hence, the present study used dialogical argumentation-based instruction (DAI) to explore how students develop and revise their argumentation strategies while solving various kinds of ill-defined problems. The study is guided by the following research question: How do students exposed to dialogical argumentation-based instruction develop and revise their argumentation strategies while solving science IDPs in groups?

(Iwuanyanwu, 2022)

Many studies related to this issue have been conducted over the past decades, addressing various aspects of interest in S&T and progressively building knowledge in this field. Analysis of these studies and related syntheses (Krapp & Prenzel, 2011; Potvin & Hasni, 2014; Renninger & Hidi, 2011; Schraw & Lehman, 2001) shows that while a lot has been learned about interest, further research is still needed, particularly in different cultural and educational contexts—since interest seems to depend on these contexts (Ainley & Ainley, 2011; Wang & Berlin, 2010) — with a focus on classroom teaching methods (House, 2009; Palmer, 2009). Osborne, Simon, & Collins (2003), in a review on attitude toward S&T, note that "there is a greater need for research to identify those aspects of science teaching that make school science engaging for pupils" (p. 1049). Other authors have also pointed out the need to develop research and tools that simultaneously take into account a number of interest-related components (Lamb, Annetta, Meldrum & Vallett, 2012). The contribution of this article is precisely to address these preoccupations, as indicated by the research objectives below.

(Hasni et Potvin, 2015)

While the problem of reasoning with erroneous information has been studied by researchers in cognitive psychology (Halpern, 2002; Kahneman, 2011; Sperber et al., 2010), research on pedagogical methods to address this issue is still emergent. As such, it is necessary to develop and assess the efficacy of the strategies that students need to improve their ability to assess claims about scientific issues.

The present study tested an intervention targeting the evaluation and critique of flawed scientific claims in a Web-based media context. Specifically, we investigated the question, "Does prompting for critique heighten students' ability to be vigilant of flaws in scientific information?"

(Tseng et al., 2021)

Despite the remarkable number of studies documenting the benefits of students learning collaboratively at the primary and secondary levels of education, the body of literature that has identified the relation between collaborative learning activities and college student outcomes is not quite as developed. Further, the evidence specifically linking collaborative learning strategies to college students' openness to diversity is exceedingly small.

(Loes et al., 2018)

Challenges and Benefits of Outdoor Science Education for Students' Learning

Many articles have identified challenges to introducing outdoor science in formal educational contexts, such as national assessments that do not require the use of outdoor learning environments (Dillon et al., 2006; Fisher, 2001), a lack of teacher expertise in teaching outdoors (Lustick, 2009; Skamp & Bergmann, 2001), and unpredictable weather (Dyment, 2005; Glackin & Jones, 2012). In most recent publications in the field of outdoor science education in schools' immediate surroundings, the research questions focus less on identifying challenges and more on studying the benefits of concrete outdoor pedagogical interventions.

The increased research over the last decade on outdoor science in schools' immediate surroundings reflects a desire to study a learning environment that is generally underused in schools but can contribute to the achievement of science learning objectives. One of the most frequently mentioned benefits is that outdoor science offers the opportunity to contextualize scientific concepts in authentic settings (e.g., Fägerstam & Blom, 2013; Lustick, 2009), which allows "their relevance to become immediately obvious" (Sahrakhiz, Harring, & Witte, 2018, p. 223). In the outdoors, students can also develop scientific field skills (Glackin, 2016; Glowinski & Bayrhuber, (2011); Glowinski & Bayrhuber, 2011 that might not necessarily be developed in a classroom or laboratory (James & Williams, 2017; Lavie Alon & Tal, 2017). Moreover, schools' immediate surroundings provide various environments for developing competencies in deploying science learning in new contexts, that is, transferring students' learning from one situation to another (Chen & Cowie, 2013; Glackin, (2016). Overall, the scientific work in recent years has demonstrated that outdoor environments are more than ordinary learning boosters; they are rich contexts that can lead to quality, meaningful, and authentic science learning for many students.

The main outcomes that have been investigated in scientific articles examining the benefits of outdoor science education include: (1) learning related to ecology (e.g., Ben-Zvi Assaraf & Orion, 2009; Fisher-Maltese & Zimmerman, 2015) or environmental education (e.g., Carrier, 2009; Hyseni Spahiu, Korca, & Lindemann-Matthies, 2014), (2) development of students' attitudes/ motivations/interest (e.g. Bølling, Hartmeyer, & Bentsen, 2019; Dettweiler, Lauterbach, Becker, & Simon, 2017) Becker, & Simon, 2017, (3) teachers' positive perceptions regarding outdoor learning (e.g. Borsos, Patocskai, & Boric, 2018; Glackin, 2016)2016, and (4) students' positive perceptions of outdoor learning (e.g., Carrier, Thomson, Tugurian, & Stevenson, 2014; Dhanapal & Lim, 2013). However, in a previous meta-synthesis, we concluded that "students do not necessarily perceive a clear connection between the outdoor learning they perform and its scientific value" (Ayotte-Beaudet, Potvin, Lapierre, & Glackin, 2017, p. 5351). The research we present in this article therefore aims to shed light on the benefits of outdoor science education for middle-school students by studying their general perceptions of learning in their schools' immediate surroundings.

(Ayotte-Beaudet et Potvin, 2020)

Teachers' perceptions of NOS

Research into perceptions and beliefs of NOS has contributed to the body of knowledge around how teachers understand NOS, as well as their pedagogical approaches to teach NOS (Abd-El-Khalick & Lederman, 2000). Some comparative studies have compared perceptions of NOS between teachers and their students (e.g. Dogan & Abd-El- Khalick, 2008), between pre-service and practising teachers (e.g. Hoh, 2013; Tairab, 2001), or across teachers with different cultural backgrounds such as the United States & Nigeria (e.g. Akerson et al., 2000; BouJaoude et al., 2011). Furthermore, attempts have also been made towards changing and improving those views (Akerson et al., 2000; Hansson & Leden, 2016). Overall, the consensus of the work on teachers' perceptions of NOS is that teachers, generally, lack an adequate understanding about the NOS (BouJaoude et al., 2017; Tairab, 2001; Tsai, 2002; Yacoubian & Khishfe, 2018). Furthermore, there is evidence of somewhat naive views and misconceptions about the NOS (Irez, 2006; Lederman et al., 2019).

For instance, Tsai (2002) carried out a study to explore middle and high school science teacher's (n = 37) perceptions of science learning, teaching, and NOS. Findings from individual interviews revealed that more than half the participants held traditional beliefs that are not in line with NOS. In a similar vein, Al-Omary (2006) investigated science teachers' (n = 17) beliefs on the NOS, the learning and teaching of science, and how these might be related to their teaching practices. Data collected through individual interviews alongside the video-recording of classes found that 35% of the teachers held traditional beliefs about NOS, 24% held informed beliefs, and 41% held a mixture of both.

Likewise, Saleh and Khine (2014) conducted a study to assess UAE pre-service teacher education students (n = 24) views of NOS using an open-ended instrument, namely the Views on Nature of Science Questionnaire Form C (VNOS-C) adapted from Lederman et al. (2002). This instrument particularly focused on the following aspects of the NOS: the empirical nature of science; the relationship between observation, inference, and theoretical entities; the way how theories and laws are different; creativity and imagination of scientific knowledge; the theory-laden aspect of scientific knowledge; the cultural and social nature of scientific knowledge; the myth of the scientific method; and the tentative aspect of scientific knowledge. Findings from this study indicated that many of the NOS reported views were either ill-informed or ambiguous, such as, the definition of science as a study, the experimental nature of scientific knowledge, and the distinction between a scientific theory and a scientific law, amongst others. In addition, Tairab (2001) investigated how aspiring teachers and in-service teachers perceive NOS using a developed questionnaire that addresses NOS elements such as scientific knowledge and the relation between science and society. The findings reflected having an understanding of science and research that is mainly oriented either by content or process, as well as reported having adequate views about scientific knowledge in terms of tentativeness and the description of theories, laws, and facts.

Furthermore, many correlational studies attempted to relate teacher's understanding and successful teaching of NOS to certain potentially determining factors, such as grade point average, experience, academic major, courses, training, and subject matter knowledge (Azninda & Sunarti, 2021; Bell et al., 1998; Gheith & Aljaberi, 2017; Hoh, 2013). Other studies suggest that NOS views may be attributed to varying demographic variables such as intellectual level (Akerson & Buzzelli, 2007), background culture, and religious beliefs (Dogan & Abd-El-Khalick, 2008). There is, however, a need for more research that investigates in what way the difference in attainment of NOS perceptions may be correlated with such variables.

(Schofield et al., 2023)

Serious games are generally defined as digital software the primary purpose of which is learning rather than entertainment (Klopfer, Osterweil, & Salen, 2009). They could be a beneficial alternative to other instructional methods (Griffiths, 2002; Munienge & Muhandji, 2012; Scanlon, Morris, Di Paolo, & Cooper, 2002) and could transform education (Shaffer, Squire, Halverson, & Gee, 2004) because: (1) simulation and video games let players participate in worlds otherwise inaccessible to them and thus develop new situated understanding; (2) video games make it possible for players to participate in very large scale communities of practice and to learn by doing the ways of thinking that organise those practices. With regard to the role of serious games in education, some scholars (e.g. Prensky, 2001) have even argued that developing digital-based educational games is a 'moral imperative', as learners of the new generation do not respond as effectively to more conventional instruction. In a widely publicised report following its 2006 Summit on educational games, the Federation of American Scientists (FAS) echoed these opinions by writing that 'Given the digital natives' affinity for digital technologies, digital games for learning could be potentially powerful tools for teaching' (p. 17). Of course, one must recognise the complexity of the 'digital natives' notion and raise questions about the empirical evidence supporting it (Helsper & Eynon, 2010), consider the social inequalities related to Internet and digital technology access (Camerini, Schulz, & Jeannet, 2018), recognise the complexity of changing policies and practices (Coburn, 2004), and the profound challenges related to implementing constructivist instruction (Windschitl, 2002), but there still appears to be enough potential to justify at least some private- and government-sector investment in educational game research.

Despite the potential and the popularity of serious games, there is currently no consensus with regard to their impact on science learning. At times, empirical evidence supports higher science learning achieved by students subjected to serious games in comparison with more conventional instructional methods (e.g. Huppert, Lomask, & Lazarowitz, 2002; Kollöffel & de Jong, 2013; Myneni, Narayanan, Rebello, Rouinfar, & Pumtambekar, 2013; Pyatt & Sims, 2007, 2012; Cameron, 2003; Zacharia, Olympiou, & Papaevripidou, 2008). In this context, more conventional instruction can be considered the opposite of a more active approach (previously defined) and refers mostly to lectures, discussions, textbook readings, exercises and problem solving (McLaren & Kenny, 2015; Waldrop, 2015) or, more generally, of less involvement and fewer higher-order thinking tasks. Of course, one has to acknowledge that the opposition between active and more conventional approaches can be misleading because quality of both active games and conventional instruction have changed over the years. At other times, no difference in science learning achievement is found between serious games in comparison with more conventional instructional methods (Zacharia & Olympiou, 2011; Corter, Esche, Chassapis, Ma, & Nickerson, 2011; Lang, 2012; Renken & Nunez, 2013; Wiesner & Lan, 2004). The National Research Council (National Research Council, 2011, p. 54) concluded that 'Evidence for the effectiveness of games for supporting science learning is emerging, but is currently inconclusive.' They observed that, even if the research on simulations is stronger than the research on games, both have not yet been studied enough to reach a definitive conclusion (Clark et al., 2009 & Sugrue 1988). This can be explained partially by the rapid changes in technology and the related difficulty in focusing the research. Another problem is the poor or missing description of the variables describing the context or the students that could also influence learning. Some methodological issues were also identified, such as small sample size, ecological biases related to nested groups, wide range of theoretical perspectives and variability of instruments to measure learning outcomes.

(Riopel et al., 2019)

Indeed, in recent years, a growing number of studies have argued that many frequent non-scientific conceptions (sometimes designated as "misconceptions") will not vanish or be recycled during learning, but will on the contrary survive or persist in learners' minds even though these learners eventually become able to produce scientifically correct answers. Shtulman & Valcarel (2012), for example, measured the reaction times of 150 college undergraduates who were asked to judge if scientific and unscientific sentences, taken from 10 domains of science and mathematics, were true or false. Some of these statements were labelled as "consistent" when they were usually both true or false both from the novice's and expert's standpoints (e.g. "rocks are composed of matter"; "numbers are composed of matter") and "inconsistent" when they were more frequently false according to experts, but frequently true for novices or vice versa (e.g. "fire is composed of matter"; "air is composed of matter"). Results for correct answers showed that "participants were significantly slower at verifying inconsistent statements than at verifying consistent ones, both across domains [. . .] and within domains [. . .]" (Shtulman & Valcarel, 2012, p. 212). Results were robust to differences in true-falseness of statements. Also, these authors argued that results could neither be attributed to differences in the syntactical or linguistic forms of the sentences nor to the level of familiarity with the information given. It was therefore suggested that if some correct answers required more time to be produced than others, it was most likely because they required more demanding cognitive processes. Since differences in reaction times matched conditions that differed in the involvement of common misconceptions, lags were attributed to the suppression of these conceptions.

Results that support this interpretation were obtained by Babai & Amsterdamer (2008) in conceptions about solids and liquids and by Babai, Sekal & Stavy (2010) about living things. The former research involved images of rigid, non-rigid or powder solids and of runny or dense liquids that had to be correctly classified as "solids" or "liquids". Since naive conceptions about liquids and solids often unduly involve some directly perceptible properties, like "pourability" or "hardness", the correct qualification or disqualification of some of the represented substances was presumed to be less intuitive, as in the case of Plasticine or honey. Indeed, reaction times confirmed that "reasoning processes associated with correct classification of objects that are not consistent with the naive conceptions are more demanding" (Babai et al., 2010, pp. 556–557), and they therefore argued that "naive conceptions in young children persist and affect junior high school students" (Babai et al., 2010, p. 557). The second study also used images. Pictures of living and non-living things were presented to 15- to 16-year-olds. Some of these stimuli were considered as "intuitive" (insects, mammals, static objects) and others as "less intuitive" (flowers, celestial bodies, vehicles) because of the naive idea that living objects usually move. Since correctly classifying plants took longer than classifying animals and that classifying dynamic non-living objects took longer than classifying static ones, the authors argued that "despite prior learning in biology, the intuitive conception of living things persists up to 15 to 16 years of age, affecting related reasoning processes".

(Potvin et al., 2015a)

The paper is structured into standard main parts. The theoretical framework provides an introduction to key concepts such as intrinsic motivation and situational interest. In the Research context, we introduce four research questions which we return to in the Discussion, where the results are seen from the perspective of previous research and with possible implications for the practice of physics teaching and learning (with a focus on hands-on practical work).

(Kácovský et al., 2023)

This article is organized as follows. The Methodology section discusses the measurement of emotional engagement and describes the study design, including the sample, protocol, and experimental processes. The Results and discussion section presents the findings of this study, relating them to the hypothesis and questions raised in the Introduction section. The Conclusion section summarizes the major findings of this study, highlights their importance, and discusses future areas of research.

(Allaire-Duquette et al., 2014)

Student engagement

Recognized as a complex and multifaceted construct, student engagement is rooted in action (Bond et al., 2020; Kahu, 2013; Reschly & Christenson, 2012). Here considered in a course context, it represents the investment and energy that students devote to learning (Borup et al., 2020; Fredricks et al., 2016; Skinner & Pitzer, 2012). Often described as a multidimensional psycho-social process, numerous authors (Bond et al., 2020; Christenson et al., 2012; Fredricks et al., 2019; Kahu, 2013; Lawson & Lawson, 2013; Manwaring et al., 2017; Schindler et al., 2017) refer to the definition provided by Fredricks et al. (2004) based on a qualitative literature review. These authors define student engagement as having three interrelated dimensions: behavioral, emotional, and cognitive. In a course, student behavioral engagement concerns their participation in activities and compliance with rules or norms. Next, student emotional engagement refers to their emotional reactions to activities, peers, and the teacher, and their sense of belonging to the course. Finally, student cognitive engagement corresponds to their psychological investment in activities to master complex knowledge, as well as their use of learning or metacognitive strategies. Christenson et al. (2012) summarize student engagement by stating that "engaged students do more than attend or perform academically; they also put forth effort, persist, self-regulate their behavior toward goals, challenge themselves to exceed, and enjoy challenges and learning" (p. v).

Blended learning

BL is situated on a continuum between face-to-face and online learning (Lakhal & Meyer, 2019). The Handbook of Blended Learning defined BL as a combination of face-to-face and online activities (Bonk & Graham, 2012), while some more precise definitions explicitly identify a decrease in face-to-face meetings (Bates, 2018; McGee & Reis, 2012; Picciano, 2009), e.g., between 30 and 79% of online learning (Allen & Seaman, 2016). Although this point is not necessarily made explicit in the literature, the present study considers that such a decrease should be inherent to BL so as to avoid a one-and-a-half course phenomenon (McGee & Reis, 2012).

Given the aim of BL to combine the benefits of synchronous interactions with online flexibility and considering the improvements in digital technologies, new BL environments that allow synchronous activities to happen online instead of face-to-face for all or part of the students have also emerged in the last 15 years (Lakhal et al., 2017, 2020). The literature describes three common BL environments: 'Traditional' Blended, Blended Online, and Blended Synchronous courses (Lakhal & Bélisle, 2020; Lakhal et al., 2020; McGee & Reis, 2012). First, Traditional Blended Courses combine face-toface with asynchronous online T&L activities. Second, Blended Online Courses combine synchronous and asynchronous online T&L activities (Power, 2008; Power & Vaughan, 2010). Also found in the online learning literature, they are part of BL since synchronous online meetings enable real-time interactions between students and teachers, as is the case for face-to-face meetings. Finally, Blended Synchronous Courses combine asynchronous online with synchronous face-to-face/online activities where on-campus/ remote students simultaneously participate (Bower et al., 2015; Lakhal et al., 2017, 2020; Raes et al., 2019, 2020). In some cases, students may also have the possibility to watch meeting recordings or alternative videos instead of participating in synchronous T&L activities, thus being offered full flexibility of participation corresponding to HyFlex Courses (Beatty, 2007, 2014, 2019).

(Heilporn et al., 2021)

Background

What is Cognitive Conflict?

Definitions of cognitive conflict have been proposed by several authors. Lee defines it as "a perceptual state in which one notices the discrepancy between one's cognitive structure and environment (external information) or between the components of one's cognitive structure (i.e., one's conceptions, beliefs, substructures, etc., which are part of the cognitive structure)" (Lee & Yi, 2013, p. 603). In their 2012 article, Lee and Byun proposed an interesting and quite complete review of the available definitions. Most of them encompass the ideas of awareness, disequilibrium, not-confirmed expectations or predictions, logical conflicts between conceptions, etc. (Lee& Byun, 2012, pp. 944-945).

There are different types of conflicts. Some occur between conceptions that exist within the same person, other conflicts occur between different people (socio-cognitive conflicts), but, most often, the models aim to trigger conflict by introducing new, contradictory information (Limon, 2001, p. 360). In this case, a contradiction occurs, for example, between a student's conception or expectations and the crucial information that a knowledgeable teacher brings to the student's attention and which the student perceives to be discrepant. Thus, such a "discrepant event" can be defined as "the physical experience that provides students with novel evidence to contradict their existing conceptions" (Kang, Scharmann, & Noh, 2004, p. 73). Baddock and Bucat also provide an interesting definition, suggesting that conflict is

a puzzling situation which is counter-intuitive. The discrepancy between expectation and observation is often brought into sharp focus by asking students to predict what will happen before the demonstration is conducted (2008, p. 1115).

Many different means of triggering conflict have been proposed. Among others, "baffling demonstrations or paradoxes to arouse [. . .] motivation" (Lee et al., 2003, p. 586) are recommended to teachers who "should probably not be afraid to overemphasize their examples or illustrations and evoke cases of striking differences" (Potvin, Masson, Lafortune & Cyr, 2015).

To summarize, cognitive conflicts are used by teachers to destabilize learners, force awareness, etc.; to ultimately produce positive outcome on learning. However, the nature of the effects they produce is subject to many hypotheses.

(Potvin et al., 2015b)

Developing interest in physics

Interest is a motivational variable that refers to a psychological state of engagement or predisposition to re-engage in a certain task, event, or idea (Hidi & Renninger, 2006). The affective components of interest include the expression of positive emotions and are neurobiologically based (Hidi, 2003). The human motivational system originates from emotional brain circuits (Panksepp, 2003). Therefore, the interest of individuals would be aroused through interactions in which they are engaged physically, cognitively, and symbolically with the object of their interest (Hidi & Renninger, 2006).

Second, both the affective and cognitive components of interest have biological roots (Hidi, 2003). Neuroscientific research on approach circuits in the brain (e.g. Davidson, 2000) and on seeking behavior (e.g., Panksepp, 1998, 2000) indicate that interested activity has a biological foundation in all mammals. Panksepp and his colleagues specifically argued that the seeking system is an evolutionary and genetically ingrained emotional brain system. (Hidi & Renninger, 2006, p.112)

Although individuals are capable of developing interest on their own, the object and the environment largely define the direction of the interest and the scope of its development (Hidi & Renninger, 2006). In physics education, interest is considered a psychological state that reflects the relationship between a student and the teaching approach of the phenomena of the material world (Hoffmann, 2002). As Kerger et al. (2011, p.608-609) report,

In the first phase, situational interest is triggered, for example, by environmental conditions such as group work, computers, puzzles, incongruous or surprising information, character identification (...)

The first phase of the development of interest is known as triggered situational interest. It involves a sudden change in an individual's affective process, which is regarded in this study as emotional engagement. In physics classes, situational interest can be triggered by the classroom climate, teaching strategies, learning activities, or contexts in which the physics lesson is presented (Häussler & Hoffmann, 2002). The learning environment can also generate situational interest among students who have, a priori, low levels of interest in physics (Hoffmann, 2002). Triggered situational interest in physics is seen as a three-dimensional concept (Gardner, 1985), with the dimensions being interest in a subject (e.g., optics), interest in the context in which the subject is presented (e.g., photography), and interest in the type of learning activity (e.g., building a camera). This study focuses on interest in the context in which the subject is presented.

(Allaire-Duquette et al., 2014)

Teachers' attitude toward science and teaching S&T

Although definitions of attitude vary, authors agree that attitude is a complex construct involving three dimensions (Ajzen & Fishbein, 2005; Jones & Leagon, 2015; Venturini, 2007). In S&T education, the cognitive dimension includes beliefs and convictions about science as well as perceptions about the importance of science for oneself and for society; the affective dimension refers to the feelings about science and science teaching; and the behavioral dimension refers to behavior that consistently responds favorably or unfavorably to a given scientific object (Reid, 2006).

Attitude is often measured as a single subcomponent such as beliefs and perceptions about science (Jones & Leagon, 2015; Tsai, 2002), self-efficacy (Bursal, 2012; Cantrell et al., 2003; Eshach, 2003; McKinnon & Lamberts, 2014; Riggs & Enochs, 1990), and pleasure (Kazempour, 2014; Zembylas, 2002). Measuring a single subcomponent does not account for the multidimensional nature of attitude and the dynamic interaction that exists between these subcomponents (Barmby et al., 2008; Van Aalderen-smeets et al., 2012). As part of their reflection on teachers' attitude toward science and teaching science, Van Aalderensmeets et al. (2012) proposed a new theoretical framework, largely inspired by the theory of planned behavior, which emphasizes the link between beliefs and behavior (Ajzen, 2002). Figure 1 illustrates their three-dimensional model, composed of cognitive beliefs, affective states, and perceived control.

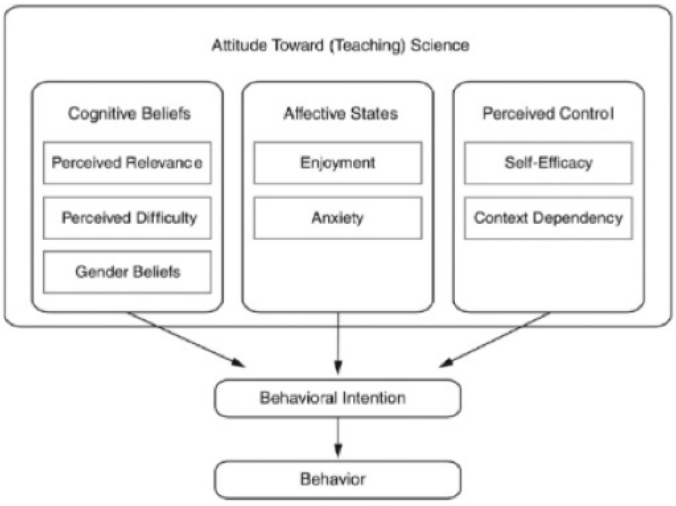



**Figure 1.** *Proposed theoretical framework for the construct of elementary teachers' attitude toward teaching science* (Van Aalderen-smeets et al., 2012, p. 176).

The cognitive beliefs dimension includes subcomponents of teachers' perceptions about the relevance of elementary S&T education, the difficulty of teaching S&T, and gender-related beliefs about students' S&T learning. The affective states dimension includes enjoyment and anxiety involved in teaching S&T. These two sub-components are not necessarily opposites and can coexist for a teacher who experiences enjoyment despite some stress during scientific activities.

(Marec et al., 2021)

The concept of interest is used in a variety of research fields, including psychology, educational psychology, sociology, S&T education (Krapp & Prenzel, 2011). Our study is located primarily in this last field. Our focus is to understand phenomena related to S&T education rather than to make a contribution to theorization about the concept of interest per se. The conceptual framework will therefore present the way we use the concept of interest to determine, based on writings in this field, the main dimensions (components) and indicators to use in order to develop tools for data collection and analysis. As Renninger & Hidi (2011) underscore, "The construction of a theoretically satisfactory interest measure requires a specification of the interest construct or a particular aspect of this construct that is used as a basis for operationalization" (p. 36). The synthesis published by these authors (Renninger & Hidi, 2011) clearly shows that there is no one stabilized and fully agreed-upon theoretical orientation towards the concept of interest. However, "general agreement can be found with regard to the central characteristics of the interest construct" (Krapp & Prenzel, 2011, p. 30). Three sets of characteristics can be found in most texts and will be used as a basis here to operationalize the concept in our research: a) the attributes of the concept of interest; b) the dimensions that make up this construct; and c) the analytical levels on which it is examined in studies.

Characterization of the concept of interest

Following on Gardner (1996, in Krapp, 2007), many authors (Hidi, Renninger & Krapp, 2004; Krapp, 2007, Krapp & Prenzel, 2011; Renninger & Hidi, 2011) consider that "the decisive criterion of the interest construct which enables it to be clearly distinguished from several neighbouring motivational concepts [such as attitude and motivation] is its content specificity" (Krapp & Prenzel, 2011, p. 30). "One cannot simply have an interest: one must be interested in something" (Gardner, 1996, p. 6, in Krapp, 2007). "The interest construct is conceptualized as a relational concept: an interest represents or describes a more or less enduring specific relationship between a person and an object in his or her life-space" (Krapp, 2007, p. 8).

The object of interest in the field of S&T can be a specific subject (biology, physics, chemistry, etc.), a specific area or field of knowledge (the study of animals), a concrete operation or object (lab manipulations), an abstract scientific activity (formulating a scientific problem or question, or analyzing data), etc. (Häussler, 1987; Häussler & Hofmann, 2000; Krapp, 2007; Krapp & Prenzel, 2011). When discussing S&T as an object of interest, it is also important to distinguish the way S&T are perceived in society (outside school) from the way it is taught and learned in school context. The focus of our research is the latter.

(Hasni et Potvin, 2015)

2.2 | Defining science misinformation

Misinformation is partially or entirely incorrect information that is distributed without the intent to deceive. Disinformation is inaccurate information shared with an intent to deceive or harm (Wardle & Derakhshan, 2017). In this study, we define scientific misinformation as information that contains scientific errors or erroneous ideas about science, but do not address intent, since in the present study it is difficult to establish intent when misinformation is communicated. We draw from social theories of ignorance by Smithson (1985), in which erroneous ideas can be characterized by incompleteness (i.e., in the form of omission or ambiguity) or distortion (i.e., in the form of mistaken information). For example, incompleteness in scientific claims may include claims taken out of the study's context; whereas distortions of science may involve biased assessments of data (Allchin, 2001). This definition of misinformation encompasses content that contains inaccuracies about scientific concepts and interpretations using flawed scientific reasoning, regardless of the author's intent.

(Tseng et al., 2021)

3.1 | "Thinking slow" to maintain epistemic vigilance against Misinformation

This study is based on the theoretical notion that people can avoid falling prey to deception by consciously enabling a critical awareness, or epistemic vigilance, of communicated messages (Kahneman, 2011; Sperber et al., 2010). Sperber et al. (2010) describe epistemic vigilance as "not the opposite of trust; the opposite of blind trust (p. 235)," and they argue that trustful communication between individuals can only be possible if humans possess cognitive mechanisms to identify what is untrue. By doing so, individuals limit potential harm of relying on testimony without question (Gierth & Bromme, 2020). Epistemic vigilance can be achieved by "thinking slow" or thinking critically to assess a communicator's perceived competence and benevolence, and to evaluate the validity of claims and justifications (Kahneman, 2011; Sperber et al., 2010). Research in cognitive psychology and philosophy has suggested consistently that humans are trustful by nature and are inclined to accept information they are presented with, especially when the information is familiar and confirms their pre-existing attitudes (Gilbert, Krull, & Malone, 1990; Mercier & Sperber, 2011; Nickerson, 1998; Petty & Wegener, 1999; Swire, Ecker, & Lewandowsky, 2017).

Additionally, in today's information landscape, competing claims are abundant and come from sources of wide-ranging quality, heightening the risk of being misinformed. As such, the act of assessing information critically is not an intuitive process, but one that requires a great deal of deliberation and effort (Halpern, 2002; Kahneman, 2011; Sperber et al., 2010). Research in cognitive science has suggested that thinking takes on two broad forms—System 1 "Fast" Thinking, which is effortless and largely based on intuition, and System 2 "Slow" Thinking, which is effortful and deliberate (Halpern, 2002; Kahneman, 2011).1 System 1 "Fast" Thinking allows for swift decisions in the absence of time and information (Gigerenzer, 2007), but at the cost of limitations to human cognition such as bounded rationality (Simon, 1978) and distortions of social and emotional biases (Halpern, 2002). In contrast, System 2 "Slow" Thinking involves the effortful task of evaluation—judgment of a source's credibility, weighing of evidence, assessment of risk, and evaluation of arguments—actions that comprise the process of thinking critically.

(Tseng et al., 2021)

Self-Concept

The general idea of self-concept usually refers to "collections of beliefs that individuals hold about themselves" (Geertje et al. 2010). For example, self-efficacy would be a particular type of self-concept that normally "refers to individuals' beliefs about their ability to successfully produce outcomes" (Renninger and Hidi 2016, p. 82). Very often, in educational studies, it is measured by asking individuals to compare themselves with other people or groups and/or with the abilities that they are assumed to have. In this context, to demonstrate a relatively high self-concept, "one must achieve more with equal effort or use less effort than do others for an equal performance". (Nicholls 1984, p. 328) Thus, and since the focus of our study is S&T, we might infer the latent variable self-concept with items such as "Compared to all other students, I consider myself to be (very weak -> very good) in S&T."

This relative perception of ability is usually considered as a rather stable trait, but which can also be influenced by the amount of experienced difficulty/easiness in resolving day-to-day tasks. It is believed to be very easy to weaken if one feels that he/she has not satisfied an authority figure's explicit expectations. Since grades allow students to compare each other, they are also naturally expected to affect it. However, even though we can easily hypothesize that achievement should be a predictor of self-concept, it is also sometimes found to be a product of it (Helmke and van Aken 1995), especially with younger students (Ganley and Lubienski 2016). Indeed, it appears reasonable to believe that having a strong self-concept encourages engagement, which in turn favors success. But Ganley & Lubienski argue that "most existing studies of the relation between confidence and achievement that use large samples and rigorous statistical techniques support a reciprocal effects model" (2016, p. 184).

A reciprocity of a similar kind has been observed in mathematics education between selfconcept (self-efficacy) and the construct of interest (Marsh et al. 2005). Whether this reciprocity is balanced is however contested. Indeed, sometimes an "interest towards self-efficacy" predominance was recorded (at the secondary level; N = 6908; grades 7–10) (Bong et al. 2015), while in other instances, the opposite was also observed (Ganley and Lubienski 2016, p. 188), although with weaker beta (β) values (0.12). We therefore agree with Bong's premise "the causal predominance between them remains contested" (Bong et al. 2015, p. 37).

Naturally, self-concept is also considered a good predictor of the intention to pursue science, although not as strong as interest. In a previously published cluster analysis, we suggested that this may merely be due to ordinary preferences. Indeed, about 25% of students appear to have a strong S&T self-concept but just lack enough interest needed to develop the intention to pursue. Called "confident indifferents," these usually performant students were hypothesized to "simply prefer other subjects over S&T" (Potvin and Hasni 2017a, p. 16).

(Potvin et al., 2018)

Interest

The general construct of interest refers to the "psychological state of a person while engaging with some type of content" (Renninger and Hidi 2016, p. 8). In the research literature, "the main references supporting the use of the interest construct [are] to Krapp and Hidi's work, which insist on the 'relationship (generally positive) between individuals and objects'" (p. 94). Interest is thus always in something and this "something" is usually indicated in the items used to assess it, such as disciplines, objects, or specific situations. Like the construct of attitude, interest is often considered to be composed of affective, cognitive, and behavioral dimensions. An additional "value" component is frequently added when assessing it (with agreement items such as "science is something important"). The most popular model of interest is the "four-phase model" (Hidi and Renninger 2006). While the initial phases of the development of interest are considered as context- or situation- dependent (situational interest) and therefore more volatile, the last ones have been described as more resilient traits (individual interest). The model suggests that repeated situationally interesting events can eventually increase the likelihood of developing an enduring and resilient individual interest. Since our research focuses on longitudinal evolution of perceptions and less on small-grain events, we will thus focus on the construct of individual interest. And since the educational focus of this research, the considered objects of interest will be science and technology as they are experienced by students in school, with agreement items like "I look forward to the next S&T lesson."

Though it could reasonably be hypothesized that interest can be found at the root of learning (like many correlational studies suggest (Singh et al. 2002)), researchers using cross-lagged panel designs have (surprisingly) often found no convincing causal relation between interest (or intrinsic motivation) and achievement (Marsh et al. 2005; Weidinger et al. 2017). Leibham et al. (2013) for example, found no evidence that early science interests (even in preschool) predicted later science achievement in 8-year-old boys, although it moderately predicted achievement in girls (Leibham et al. 2013). Some exceptions however do exist. Indeed, it was elsewhere found that "math performance predicted later interest, but interest did not predict later performance" (Ganley and Lubienski 2016, p. 189). Still, Ganley notes that the current literature "provides inconsistent evidence for this relation" (2016, p. 184).

Of course, interest is also easily considered a precursor to the intention to pursue studies and a career in S&T (Taskinen et al. 2013). In fact, most scientists and engineers say that they became interested in science at an early age (Maltese and Tai 2010). However, we believe that the reverse hypothesis is often overlooked, especially with younger students. Indeed, youngsters' intentions to pursue science may not always be motivated purely by interest. For instance, they might have family members in scientific professions, hold assumptions about the superiority of careers in S&T, be subjected to peer pressure, or simply have a positive or even exaggerated perception of scientists and of what they can do (e.g., superheroes, such as Iron Man and the Hulk, are often scientists). It is thus conceivable that, before being progressively introduced to more rigorous S&T, students might express interest for reasons that are in fact external to S&T. And it is not unreasonable to think that these intentions could be causal and resilient forces that influence their perception set. Therefore, the potentially reciprocal relationship between intentions to pursue and other perceptual constructs might be worth monitoring during these crucial years.

(Potvin et al., 2018)

Motivation

Motivation, which is a tendency to behave in a specific direction, has two main dimensions: intrinsic and extrinsic. Extrinsically motivated students seek out external rewards for their behavior in the shape of high grades, academic honors, scores on tests, and awards from parents or teachers. Intrinsically motivated students engage in learning activities to satisfy their interest or curiosity. This type of motivation reflects students' intrinsic interest in the content, materials, or task. Extrinsically motivated students also engage in learning to satisfy their needs, but their needs are for something different. Studies that investigate motivation's association with negative or positive impacts on mathematical problem solving (Alcı, Erden, & Baykal, 2008; Ozcan, 2016) report that individuals with enough intrinsic motivation are not affected by negative external factors before the learning takes place. In the research literature, motivation and self-efficacy are regarded as nested constructs (Zimmerman, 2008) and self-efficacy is seen as a motivational measure in conjunction with internal and external motivational measures (Vollmeyer & Rheinberg, 1999). Higher self-efficacy expectations can lead to an increase in motivation (Bandura, 1986; Braten et al., 2004; Liu & Koirala, 2009).

(Özcan et Eren Gümüş, 2019)

Mathematics anxiety

Mathematics anxiety is defined as feelings of tension and discomfort that might prevent someone from carrying out his or her actual capability in mathematical problems (Ashcraft, 2002) and can lead to the development of negative attitudes toward mathematics (Tooke & Leonard, 1998). These negative attitudes can prevent students from reaching their potential in terms of mathematical capability (Hannula, 2005). Pajares (1997) indicated that when people experience negative thoughts and fears about their capabilities, these negative affective reactions can further lower perceptions of that capability and result in stress, thereby potentially exacerbating poor performance and fear. The results of those studies underlining the negative effect of mathematics anxiety on mathematics achievement (e.g., Alexander & Cobb, 1984; Hackett, 1985; Hamid, Shahrill, Matzin, Mahalle, & Mundia, 2013; Jain & Dowson, 2009; Karakolidis, Pitsia, & Emvalotis, 2016; Lee, 2009; Lee & Stankov, 2013; Pitsia et al., 2017) agree that anxiety is a crucial barrier to teaching mathematics and equipping students with problem-solving skills. Sociocognitive theory suggests that individuals who do not perceive themselves to be capable of coping with threats become stressed and experience anxiety in similar or comparable circumstances (Bandura, 1977). Consequently, their functional levels become limited. Under the exact opposite circumstances as highlighted by Bandura (1989) and Lent, Lopez, Brown, and Gore (1996) the anxiety level is diminished and the individual experiences strong self-efficacy beliefs.

(Özcan et Eren Gümüş, 2019)

Self-efficacy

Self-efficacy, defined as a judgement or assessment of one's capabilities to successfully perform a particular given task (Bandura, 1977, 1997), has been highlighted as an important predictor of academic performance in general (Braten, Samuelstuen, & Stromso, 2004; Ferla, Valcke, & Cai, 2009; Liu & Koirala, 2009) and of mathematics achievement specifically (Ferla, Valcke & Cai, 2009; Pajares & Graham, 1999; Pajares & Miller, 1995). Selfefficacy is thought to influence behavior through motivational, cognitive, and affective processes. In the research literature, self-efficacy and self-concept have been presented as being interrelated. Mathematics self-concept reflects more general beliefs about competence (i.e., "I'm good at mathematics") whereas, mathematics self-efficacy refers to much more specific and situational judgments of capabilities (i.e. "I'm confident I can solve this type of two-digit subtraction problem") (Linnenbrink & Pintrich, 2003). Although it is difficult to find empirical evidence for the existence of obvious differences between these two constructs, some research (Bong & Skaalvik, 2003; Ferla et al., 2009; Lee, 2009) has highlighted the importance of self-efficacy on mathematics performance. Students with higher mathematics self-efficacy feel confident about being able to cope with difficult mathematical problems and are more accurate in mathematical computations (Collins, 1982; Hoffman & Schraw, 2009). Students with high mathematics self-efficacy levels were more resilient and patient in the face of adversity, invested more effort and time in order to achieve, and participated more effectively in class (Pajares, 2002). Previous research has reported that mathematics self-efficacy is positively related to mathematical problem solving (Kramarski, 2004; Kramarski, Mevarech, & Arami, 2002) and mathematics performance (Hoffman & Spatariu, 2008; Kabiri & Kiamanesh, 2004; Liu & Koirala, 2009; Pajares & Graham, 1999). Pajares and Miller (1995) pointed out that mathematics self-efficacy was a stronger predictor of success in solving specific mathematical problems than of total mathematics performance.

(Özcan et Eren Gümüş, 2019)

Metacognition

Metacognition, which is a significant predictor of general achievement and especially mathematics performance (Desoete & Veenman, 2006), can be defined as being aware of one's cognitive process and managing it when necessary (Flavell, 1976). Metacognition is considered in terms of two parts: metacognitive knowledge and metacognitive experiences (Efklides, 2008; Flavell, 1981). Metacognitive knowledge includes knowledge of oneself, the task at hand, and the strategy for successfully completing the required task (Flavell, 1979, 1987). Metacognitive experience is "what a person is aware of and what she or he feels when coming across a task and processing the information related to it" (Efklides, 2008, p. 279). Metacognitive experiences provide feedback to the behavioral control process by monitoring the implemented strategy, determining whether it is being successful, and assessing the outcomes (Moores, Chang, & Smith, 2006).

When students are engaged in challenging tasks like mathematical problem solving, metacognition becomes more important (Holton & Clarke, 2015). Studies report that there is a high relationship between metacognitive skills and problem-solving skills (e.g., Jaafar & Ayub, 2010; Ozsoy, 2011). Metacognition has a complicated relationship not only with performance, but also with behavior, in that it triggers the problem-solving behavior, monitors performance, and changes behavior if things are not going as expected. There is a strong positive relationship between self-efficacy and metacognition (Cera, Mancini, & Antonietti, 2013; Hoffman & Spatariu, 2008), both of which are closely related to mathematics performance and share certain properties (Moores et al., 2006).

(Özcan et Eren Gümüş, 2019)

Collaborative and cooperative learning

Collaborative learning is thought to influence intellectual growth by requiring students to assume individual responsibility though interdependent work with others in achieving shared educational goals. The change that happens as a result of learning collaboratively occurs as a consequence of the sociocognitive conflict and attendant cognitive disequilibrium that arise in group work. Disequilibrium occurs when group members are confronted with the diversity of others' perspectives in the group (Davidson & Worsham, 1992; Piaget, 1950; Vygotsky, 1978). As group members experience with these new perspectives, they "rehearse and restructure information to retain it in memory and incorporate it into existing cognitive structures" (Johnson & Johnson, 2002, p. 120). Nelson (1994) added that student misunderstandings of new ideas and concepts may inhibit their ability to learn effectively. The diversity in perspectives associated with collaborative learning, however, allows students to identify and correct those misunderstandings, thereby enhancing the potential for student achievement.

It is important to note that substantial overlap exists between the terms "cooperative learning" and "collaborative learning" in the teaching and learning literature. Although both approaches involve students cocreating knowledge together (vs. passively absorbing information from an instructor), important differences distinguish these concepts. First, cooperative learning is typically used in K–12 classrooms, whereas collaborative learning generally applies to higher education settings (Barkley et al., 2014). Next, compared with collaborative learning, cooperative learning is a more structured method of teaching that requires a greater degree of intervention and direction from the instructor, as is commonly employed in primary and secondary education classrooms. Lastly, the definition of "cooperative learning" includes five required components: positive interdependence, face-to-face promotive interaction, individual accountability, development of social skills, and group processing (Johnson, Johnson, & Holubec, 1990). As such, and despite some similarities, cooperative learning and collaborative learning are indeed distinct instructional approaches (Bruffee, 1993, 1995; McInnerney & Robert, 2004; Pascarella & Terenzini, 2005). Given our focus on the effects of peer learning at the collegiate level and the variables in our data set, we explored the effects of collaborative learning in this study.

Collaborative learning techniques can be used for discussion, for problem solving, and for engaging students with writing. Common examples include think-pair-share activities, small-group discussions, and group-based case studies. Successful groups usually contain two to six students to maximize student interaction and involvement (Barkley et al., 2014). Additionally, effective group composition comes from instructor assignment, random assignment, or content-based interests; student-chosen groups tend to be homogenous and fail to achieve many of the goals of collaborative learning (Fiechtner & Davis, 2016).

(Loes et al., 2018)

Moderators

As can be noted from the preceding overview, several moderators have been empirically found to affect the relationship between playing serious games and learning outcomes. The following list describes the expected significant moderators that will be tested in the present meta-analysis, and the hypothesis regarding each moderator's effect based on the overview and theoretical postulates from the scientific literature. A dichotomist categorisation (i.e. theoretical and methodological moderators), as used by some previously discussed reviews (e.g. Clark et al., 2015; Sitzmann & Ely, 2011), will be used to classify moderators. The first four theoretical moderators categorise context (subject area, grade level, duration of intervention, activity level of the comparison group), three more theoretical moderators describe some qualities of the games (ludic content, level of realism, level of user control) and the last four moderators characterise methodology (randomisation, experimental design, year of publication, publishing status).

(Riopel et al., 2019)

Ludic content. Examining the moderator effect of ludic content consists of verifying whether the effect of serious games on learning achievement, in comparison with more conventional instruction, varies depending on how ludic a game is. The ludic content can be described as the 'entertainment value' of the game (Sitzmann & Ely, 2011), the 'enjoyment felt while playing' (Hays, 2005) or very simply, as Prensky (2001) wrote, how 'fun' the game is. Baranowski, Buday, Thompson, and Baranowski (2008) noted that 'fun' is not a concept that is well understood and that typical measures of enjoyment (or fun) have used synonyms of fun (e.g. enjoy, like, interested, pleasurable, energising). Several scholars (e.g. Goh, Ang, & Tan, 2008; Johnson, 1991; Prensky, 2001) posited that more ludic serious games promote greater learning achievement. In a 2009 meta-synthesis examining the effects of serious games on health and physical education, Papastergiou (2009, p. 608) notably concluded that 'enjoyment [. . .] seems to account for the effectiveness of the games.' In the present overview, it was pointed out that only Sitzmann and Ely's metaanalysis (2011) examined the effect of this moderator on the learning achieved playing serious games, and found learning attained with high entertainment value games (d = 0.26, 95% CI [0.11–0.41], N = 809) was not significantly higher than learning achieved with low entertainment value games (d = 0.38, 95% CI [0.31–0.45], N = 32163216). Thus, because of theoretical postulates and because of Papastergiou's meta-synthetic findings, the eighth hypothesis reads as follows:

H8: The beneficial effect of instruction using serious games on learning achievement outcomes is greater for games with higher ludic content than for games with lower ludic content.

One has to recognise here that the effects of ludic content could also have a complex relation to engagement and duration of gameplay. This testing would be beyond the scope of the present work.

(Riopel et al., 2019)

The consideration of the flipped classroom as only a simple rearrangement of activities has been criticized in the literature. For example, Bishop and Verleger (2013) pointed out that a definition of the new method with more added value would not only switch activities but would also include a series of additional tasks both inside the classroom (such as problem solving in groups, Chiang, 2017) and outside the classroom (such as answering questionnaires and performing practical exercises, Porcaro et al., 2016). The completion of questionnaires based on the students' readings outside the classroom was also previously considered by Moravec, Williams, Aguilar-Roca, and O'Dowd (2010). They established that these activities could be used by the instructor to update the lecture material based on what the students misunderstood or needed to improve. Additionally, instructors can use this information to provide students with appropriate feedback, which is known to be crucial when assessing how students learn (Roehl, Reddy, & Shannon, 2013; Elmaadaway, 2018) and as a key determinant of students' performance (Butt, 2014).

In addition to the feedback that the instructor provides the students, the literature has also identified as being important the feedback that students themselves provide about the flipped classroom method so that it can be determined whether the method has been perceived as an effective and helpful tool in learning (Frisby & Martin, 2010). In terms of the fourth benefit of our research, we empirically test and include as an additional dimension of our flipped classroom proposal the feedback that students directly report to the instructor with key information about the elements of the didactic material previously delivered by the instructor that needs further explanation. This two-way feedback strategy combines the students' and teacher's work and enhances the effect of the flipped classroom by providing an effective link between out-of-class and in-class activities. In defining the flipped classroom, several authors have also emphasized the role of technology, the interactive use of which is considered to be crucial in the process of moving the lecture outside the classroom and conducting more practical activities inside the classroom (Elmaadaway, 2018; Huang & Lin, 2017; Wang, Jou, Lv, & Huang, 2018). Technology allows for more flexible and studentcentered education (Eid & Al-Jabri, 2016; Wanner & Palmer, 2015). However, when discussing technology, the previous literature has mainly referred to online videos. The fifth benefit of the present research comprises the use (and empirical testing) of such online videos in conjunction with mobile devices, social networks, and cloud computing applications during the entire flipped classroom teaching-learning process.

(Murillo-Zamorano et al., 2019)

Innovative intervention to improve teachers' attitude

Studies have shown that the contents of in-service training are too often based solely on the acquisition of disciplinary content (Jarvis & Pell, 2004; Morrell & Carroll, 2003; Palmer, 2006) and that this type of training has little impact on the quality or the frequency of teaching (Diamond et al., 2014).

Self-efficacy would only improve if training programs combine both the learning of scientific concepts and the manipulation of materials in an investigative process (Bleicher & Lindgren, 2005; Johnston, 2003; Posnanski, 2002). In addition, program length seems to be important. To produce the expected results, training should be conducted over several weeks (Roberts et al., 2001) or from 40 to 100 hours (Lumpe et al., 2012).

In our literature review, we noted an apparently effective but poorly documented type of intervention to improve teachers' attitude: the pairing of resource persons (e.g., scientists or pre-university/university science students) with elementary school teachers. A study by Delclaux and Saltiel (2013) paired university students (resource persons) with elementary school teachers. In this innovative form of partnership, pairing a university student with a teacher introduces a new approach and a way to provide teachers with the tools they need to teach S&T. The university student is present in the classroom during activities conducted by the teacher and assists the teacher by co-facilitating scientific activities and answering pupils' questions if required. The authors did not seek to study teachers' attitude; they observed the implementation and adoption of hands-on activities by the teachers. Nevertheless, the authors noted that the presence of university students in elementary school classrooms enhanced S&T learning and fostered a positive attitude toward science for both the teachers and their pupils. This practice seems to have beneficial effects on the attitude of teachers, who say they feel supported in conducting scientific activities in their classroom.

At the heart of our approach was the intervention we developed, which, inspired by Delclaux and Saltiel (2013), aimed to support elementary school teachers by pairing them with pre-university science students. Included in the intervention were experimental activities to be carried out in the classroom using the scientific investigation approach. This approach, introduced in S&T programs by Quebec's Ministere de l'Education, du Loisir et du Sport (MELS) (2001), is characterized by finding solutions to scientific or technological problems. Using this approach, our intervention also addressed the teachers' learning about the nature of science (NOS). NOS refers to the epistemology and sociology of science (Lederman et al., 2001), and its teaching is strongly recommended by many authors, even at the elementary level (Lederman & Lederman, 2014; Liu & Lederman, 2007; Osborne et al., 2003). Studies conducted with future or practicing teachers have shown that expressly teaching NOS is effective in fostering better understanding of science (Abd-El-Khalick, 2001; Abell et al., 2001; Akerson et al., 2000).

(Marec et al., 2021)

Negative Priming as an Additional Indication of Inhibition

Finally, another growing argument that inhibition might be involved in the production of some correct answers is based on negative priming studies. Negative priming is recorded latency that occurs when a trial (probe) is immediately preceded by a trial (prime) in which a distractor (like a misconception) has to be actively ignored. "When primes received intentional processing, they facilitated processing of identical probes; when the same primes were (actively) ignored, processing of subsequent probes was delayed (negative priming)" (Tipper, 1985, p. 586). Even though some alternative causal explanations for such latencies—sometimes based on memory retrieval—have been proposed (Egner & Hirsch, 2005), most of the authors attribute the phenomenon to the presence of the function of inhibition (Borst, Poirel, Pineau & Cassoti, 2012; Tipper, 2001). For example, using a task where an intuitive interference (misconception) according to which "the shape with the larger area has a larger perimeter", Babai et al. (2012) tested 51 11th and 12th graders to "explore whether correctly answering an incongruent condition prime in the first part of the task would increase the response time of a subsequent congruent probe trial" (pp. 766–767). Their analyses of correct answers led them to conclude that "pre-activation of control mechanisms (inhibition of intuitive interference on an earlier problem) can interfere with a subsequent problem that is in line with intuitive reasoning". Therefore, since negative priming has in many cases been considered as an indication of the presence of inhibition, we will also investigate for this phenomenon in order to provide stronger evidence for the claim of persistence.

(Potvin et al., 2015a)

Teachers' strategies to foster student engagement in blended learning

Student engagement being malleable through pedagogy, it can be influenced by teachers' strategies, i.e., what teachers do to encourage student engagement in their courses (Fredricks et al., 2004, 2019; Kahu, 2013; Lawson & Lawson, 2013). In BL environments, there are however few studies investigating what teachers do and why they do it (Smith & Hill, 2019; Taylor et al., 2019; Torrisi-Steele & Drew, 2013), and even fewer investigating how they foster student engagement (Halverson & Graham, 2019; Halverson et al., 2014; Jeffrey et al., 2014; Manwaring et al., 2017; Siemens et al., 2015; Taylor et al., 2018). For instance, although the literature stipulates that synchronous and asynchronous modes of BL must be thoughtfully integrated in order to optimize student engagement (e.g., Garrison & Vaughan, 2008; McGee & Reis, 2012), several authors mentioned that there are few concrete recommendations in this regard (Graham et al., 2014; Manwaring et al., 2017; Siemens et al., 2015; Taylor et al., 2018). The sparse literature concerning how teachers foster student engagement in BL addresses the issue with varying degrees of specificity, ranging from specific digital technology applications to activities to general strategies.

The following literature review focuses on teachers' perspectives relating to strategies fostering student engagement in BL. Excluding course case studies, Vaughan (2014) studied the role of online collaborative learning applications to foster student engagement and success in traditional BL courses for 1st-year undergraduates. By way of a mixed methodology notably involving teachers' interviews (n = 8), the author concluded that such applications enhanced student engagement. However, he suggested that future research should explore whether the use of digital technologies stimulates student engagement in BL in and of itself or instead mediates a more general strategy such as active and collaborative learning. Montgomery et al. (2015) also examined the role of digital technologies in fostering student engagement in three traditional BL courses for education undergraduates, through teachers' narratives (n = 3). The teachers reported that students were first engaged asynchronously online using varied resources (e.g., texts, videos), choices being provided to foster student engagement. In subsequent synchronous meetings, student engagement was sustained through active learning, sometimes using an experiential approach. Then student engagement was reinforced online by individual or collaborative projects and digital resources devoted to deepening understanding of the contents. Finally, the authors stressed the importance of student-content interactions to promote student engagement. Although this publication described several teachers' strategies fostering student engagement in BL, a detailed analysis of strategies shared by the three courses was not provided and the number of courses was limited, which could be interpreted as methodological limitations.

(Heilporn et al., 2021)

A number of variables have been proposed to account for this gender imbalance, including traditional socialization practices that reinforce STEM fields as male domains, and women's lower confidence ratings in STEM (Wang & Degol, 2017). An important determinant of the gender gap in STEM fields concerns the stereotype threat that women entertain about this field. Many women believe that they have little natural ability in programming (Beyer, 2014) and their perception of self-efficacy is often low (Askar & Davenport, 2009; Michie & Debra, 2006). Self-efficacy refers to people's beliefs in their capabilities to produce given attainments and it has been shown to be a powerful predictor of one's willingness to engage and be successful in different areas of life (Bandura, 1994). Individuals with high levels of self-efficacy approach difficult tasks as challenges to be mastered rather than threats to be avoided. They set more difficult goals for themselves, exert more active self-regulation, expend more effort, persist for longer with challenges, and show resilience in the face of adversity (Bouffard et al., 2005; Klassen & Usher, 2010; Valentine et al., 2004). It is through these mechanisms that selfefficacy enhances achievements, which in turn increases self-efficacy (Bandura, 1986).

(Allaire-Duquette et al., 2022)

2.2. Effects of knowledge on skills and engagement

Research has suggested the existence of a clear relationship in which knowledge influences skills. Perfetti (2013, p. 33) established that, if a student must acquire reading competency, it is necessary for him or her to already possess the phonemic knowledge that will help to develop the ability to read. Root and Ngampornchai (2013) described the competencies acquired by students when they spend a short time abroad. They noted that students who were familiar with the language of the country of destination developed the ability to become even more fluent in the language. However, students without previous knowledge of the language developed other types of abilities to be able to communicate in the different language. Hasan (2017) pointed out that there is a significant influence of knowledge on both students' listening comprehension and their reading comprehension. Murillo-Zamorano and Montanero (2018) analyzed the competence of a sample of economics and business studies students in orally presenting academic content. Students' knowledge in terms of their capacity to provide peer feedback, as part of a peer assessment process, was confirmed as being effective in improving oral presentation skills. These findings indicate that the existence of appropriate knowledge does indeed influence students' skills. Thus, we hypothesize that:

H4. Students' knowledge directly and positively affects students' skills.

The literature has also explored the relationship between skills and engagement. There has been much research focused on the impact that the skills developed by students has on their level of engagement. For instance, Wang and Holcombe (2010) noted that certain skills, such as the students' sense of autonomy in school, positively contribute to a series of outcomes, one of which is classroom engagement. Kahu (2013) established a series of antecedents and consequences of student engagement. Among these antecedents, students' skills were considered to be among the aspects influencing how engaged that they were. More recent research by Fredricks et al. (2016) and Shernoff et al. (2016) also examined the association between these two variables. Fredricks et al. (2016) conducted a series of interviews with students and teachers to determine their perceptions of engagement in relation to math and science. Some of their findings indicated that, according to both the students and the teachers, skills such as understanding different perspectives and the ability to follow others' ideas or to explain class content to their peers are indicators of students' engagement. The students also mentioned that working with their peers influences their levels of engagement. Shernoff et al. (2016) reviewed a set of studies analyzing the association between engagement and certain skills, finding that such aspects as the existence of encouraging types of relationships among students or the students' ability to be autonomous yield a higher level of engagement. Connolly and McGuinness (2018), chap. 7 explored the digital literacy skills of young people in the European context, investigating where and how digital skills can support the inclusion, engagement and participation of young people in the digital world. In line with their study, Hong et al. (2018) suggested that is necessary to provide further efforts to understand the influences of digital skills on students' engagement. In accordance with this line of reasoning, we state the following hypothesis:
H5. Students' skills directly and positively influence students' engagement.

(Murillo-Zamorano et al., 2019)

Exogenous Factors that Are Known to Impact Perceptions

Besides influencing each other, perceptual constructs can be influenced by other pedagogically or socially relevant factors. Their possible number is infinite. However, it is possible to suggest, based on the literature, that some may have more predictive power than others. The following descriptions of these factors will be included in our design as exogenous variables.

Among them, student's grades (achievement) as well as perceived easiness (considered here as the mere opposite of the construct of difficulty) have already been discussed as possible predictors of self-concept. BThere is currently no widely accepted definition of perceived difficulty, but it has been described as being experienced when a decision maker finds it difficult to choose a certain course of action, or when it is unclear which course of action best meets a decision makers' goals^ (McCleary et al. 2014, p. 2). Therefore, the two factors (achievement and easiness) will be included in the panel design as exogenous variables.

The third factor, gender, is one of the most frequently studied in the science education literature (Potvin and Hasni 2014a). Easy to collect and to include in statistical analyses, this factor may also be popular among researchers because it raises important social justice issues, such as the popular question of whether "school science is sexist."
While major gender differences have been observed at the university level (Hango 2013), especially in physics, chemistry, computer science, and engineering, they can also be noted earlier, while not as strongly. While being equally competent, boys usually express a stronger interest (Leibham et al. 2013) and self-concept in S&T (and mathematics) than girls (Simpson and Oliver 1985), and the decline in girls' attitudes and interest is steeper during early secondary years (George 2006; Logan and Skamp 2008; Potvin and Hasni 2014a, b). These recorded differences are however usually very small (Bong et al. 2015), especially when the questions do not differentiate between fields (physics, biology, etc.) (Hasni and Potvin 2015b). However, in qualitative research (Logan and Skamp 2008) or when different activities or finer-grained topics are evoked (mycology, human-plant interaction), the difference appears to be more salient and are not always in favor of boys (Baram-Tsabari and Yarden 2007). In most structural equation modeling (SEMs) designs, gender however usually does not often manifest as a strongly predictive variable of perceptual S&T constructs (Ganley and Lubienski 2016; Garon-Carrier et al. 2016). Since the gender variable is relatively stable and predictable, it is often included in cross-lagged panel designs as a control variable (Liu and Hou 2017). For these reasons and for simple caution, it appears that gender should be included also in our study.

(Potvin et al., 2018)

Research questions

There are few studies specifically addressing student engagement in BL, and even fewer concerning teachers' strategies to foster student engagement in such environments. While studies presented in the previous section have paved the way for research about this subject, the examination of teachers' strategies in these studies was limited by the number of teachers included and their specific focus (e.g., on digital technology applications) or context (e.g., business faculty). Teachers' strategies were also studied with varying degrees of specificity, which emphasizes the need to classify these in a clear and organized way. Moreover, most publications did not present an explicit definition of student engagement and only Heilporn and Lakhal (2020) investigated the issue using a multidimensional perspective. Furthermore, almost all studies concerned undergraduate courses, and all were situated in traditional BL courses. In blended online or blended synchronous courses, in particular, the question of how to optimize student engagement is still open (Raes et al., 2020). As a result, the way in which teachers foster student engagement in BL environments (traditional blended, blended online, or blended synchronous) has yet to be studied (Graham, 2019; Raes et al., 2020; Siemens et al., 2015). Hence, the following general research question was addressed in this study: What strategies do teachers use to foster student behavioral, emotional and cognitive engagement in BL?

(Heilporn et al., 2021)

Accordingly, this study aims to fill this gap in the literature. The main aim of the study is to investigate the effects of AR technology on middle school students' achievements and attitudes towards the course, and to determine their attitudes towards AR applications. In parallel with the aim of this study, the following research questions will be answered.
(1) Is there a significant difference between the academic achievement of middle school students using AR applications and those using traditional methods?
(2) Is there a significant difference between the attitudes of middle school students towards their science course based on whether they learned via AR technology or traditional methods?
(3) What attitudes do students who used AR technology have towards AR applications?
(4) Is there a correlation between middle school students' attitude towards AR, their attitude towards the course and their academic achievement in experimental group?

(Sahin et Yilmaz, 2020)

Research Problem and Question

Considering that (a) current research about schools' immediate surroundings appears to focus on studying the benefits of outdoor concrete pedagogical interventions, rather than their challenges (or shortcomings), and profiting from this; (b) it is important to better understand students' perceptions of science learning in this educational context in order to maximize its potential; and (c) we found no empirical studies in scientific journals that focus on factors that might be related to students' perception of learning in our research context, we ask the following research question: What factors are most related to students' perceptions of learning during outdoor science lessons occurring in their schools' immediate surroundings?

(Ayotte-Beaudet et Potvin, 2020)

Objectives of the study

Analysis of studies and syntheses that we have just cited shows that interest is strongly associated with certain variables and dimensions. For example, the literature review of Potvin & Hasni (2014) based upon 228 research articles indicates that in addition to other variables (gender, grade level, country of origin, etc.), the studies highlight the important role of school-related variables (including teaching methods), self-efficacy, and sociological variables (including the socioeconomic level of parents as well as family background). The literature review of Krapp & Prenzel (2011) stresses the need to pay special attention to specific domains or scientific disciplines (biology, physics, chemistry, etc.) and to consider a comparison of S&T with the other subjects that make up the curriculum. Moreover, the same review also highlights the importance of variables such as gender, grade level, and self-efficacy.

On one hand, the selection of variables and dimensions to consider in our study stems from analysis of the earlier studies and literature reviews that we have cited above. However, although these variables and dimensions have often been examined in an isolated way in earlier research, our aim is to consider them simultaneously in one study, while seeking to establish relationships between them and general interest in S&T. On the other hand, the selection of variables and dimensions also follows the priorities jointly established by the researchers and partners of the Chaire de recherche sur l'intérêt des jeunes à l'égard des sciences et de la technologie (CRIJEST) under which this survey was conducted. The Chair is managed in partnership by two universities and six school boards (administrative entities responsible for local management of the education system) that account for more than half of the schools in Quebec (French Canadian province). The Chair's research and interventionspriorities are jointly defined through an Executive Committee composed of representatives of Chair partners. In the context of this committee, it was agreed to study, among all the variables and the dimensions reported in the literature, primarily those on which the school can act (teaching methods, self-efficacy, family involvement in cultural activities, support for students, etc.). This rationale also explains why some variables and dimensions, such as the role of the socio-economic status of parents, were not retained. It is important to note that in addition to conducting research, the Chair, based on its findings, also aims to suggest activities and interventions that target teachers, students and parents with a view to improving S&T learning and interest in S&T and related careers. Examples of these activities are presented on the website of the Chair (http://crijest.org/).

(Hasni et Potvin, 2015)

Experimental Aim and Hypotheses

In the research presented here, we explore how a science museum's short programming workshop can impact pupils' mastery experiences, an important source of self-efficacy beliefs, with a specific focus on girls.

Hypothesis 1 (H1)

Women often believe they have little natural ability in programming (Beyer, 2014) and generally tend to have low self-efficacy beliefs for programming. Thus, we expect that girls will generally report lower self-efficacy than boys before the programming workshop.

Hypothesis 2 (H2)

H2.1
Mastery experiences are acknowledged as an important factor in the development of self-efficacy (Bandura, 1994). Pupils' participation in hands-on programming activities should offer them opportunities to control and manipulate the user interface, which is crucial to developing programming skills. Therefore, we predict that on average, pupils' self-efficacy for programming will be higher after the workshop.

H2.2
Boys generally have more positive experiences with STEM activities than girls, irrespective of experimental condition (e.g. Alexander et al., 2012; Kessels & Hannover, 2008; Hoffmann, 2002). Accordingly, we predict that boys' self-efficacy will remain higher than girls' self-efficacy beliefs following the programming workshop.

(Allaire-Duquette et al., 2022)

Statement of the problem and research question

The literature examining the misconceptions that prevail or interfere in falling bodies problems is extensive. However, it remains fragmented and few research initiatives have attempted to provide descriptions of conceptual evolutions over long periods. We thus believe that we need and extended portrait across the lifespan of the conceptual challenges that teaching about falling bodies poses to educators. The research question thus becomes: What is the evolution of the prevalence of- and the interference by- conceptual attractors in falling bodies problems, as a function of schooling and of teaching at different levels? We formulate the hypotheses that the perceived cognitive utility of misconceptual resources will decrease with schooling (and teaching) and that the adherence to each scientific conception will increase. This analysis is an original contribution because it provides, with one single task, not only an account of correct or incorrect answers, but also of the adherence (and "virulence") to many conceptual attractors (pluralist account). It is also original because it provides for the first time (to our knowledge) an examination of the use of misconceptual resources in falling bodies problems that happen withas well as without- atmospheric contexts, despite the important number of studies about such misconceptions.

The investigation has been facilitated as physics (mechanics) has been taught and learned at the same levels in the province of Québec (Canada) since at least the 1960s. Indeed, in the last 63 years, mechanics and falling bodies problems have been taught almost exclusively in secondary school as an optional course ( 5th year—around age 16), at the college level for "natural sciences" students only (pre-university college [CEGEP]—around ages 17–19); and at the university level (for physics students as well as for preservice high school science teachers— around ages 20–24). Each of these courses had the precedent as prerequisite. Thus, all students who succeeded in at least one mechanics class at university had also succeeded beforehand in college and, previously, in secondary- level mechanics courses.

(Potvin et al., 2023)

Because it can be argued that natural sciences could have a special relation to serious games and because of the current inconclusiveness of scientific literature with regard to the impact of serious games on science learning achievement, a meta-analysis was conducted in the present work to answer two research questions:

(1) What is the impact of serious games on science learning achievement when compared with more conventional instructional methods?

(2) Which moderator variables influence the relationship between playing serious games and science learning achievement? A moderator variable (i.e. hereinafter referred to as moderator) is a continuous (e.g. age, school marks) or discrete (e.g. gender, ethnicity) variable that affects the strength or direction of the relationship between an independent or predictor variable and a dependent or criterion variable (Baron & Kenny, 19861).

Answering these questions should provide the best up-to-date high-level description that characterises the significant impact of serious games on science learning. It is important to note that, because of its general stance related to the available data in the previous studies, the present work could not satisfactorily address some questions thatmight prove useful to educators or designers and be related to the effectiveness of games at the instructional level, such as: 'Why is a given game better than another one?', 'What is the best way to use a given game in school?', or more generally, 'What are the best underlying educational purposes, pedagogical models, or other forms of typologies for games?' It is also important to note that, even if the diversity and limited quality of learning outcomes in science must be acknowledged, the observed significant differences, if any, can still be interpreted at a high level with appropriate caution. The focus of the present work therefore is to determine cautiously if the previously studied games, in the specific context of natural sciences and for a short list of moderator variables, had significant impact on measured learning. It is meant to give the best possible answer with available data from literature and serve as one possibly useful starting point for future studies.

(Riopel et al., 2019)

Research Questions and Hypotheses

Considering the importance of the two first years of secondary school for later trajectories and the relative lack of knowledge we have of this crucial period; considering the uncertainties that exist about the statuses and evolution of perceptual constructs and about the strength of their reciprocal relationships; and given the lack of a clear picture of the effects of other important exogenous variables that can affect them, we asked the following questions:

(1) What is the evolution and stability of the main school S&T perceptual constructs (interest, self-concept, and intention to pursue) during the two first years of the secondary school?
(2) What are the reciprocal causal relationships that exist between these constructs? And
(3) What are the effects of perceived easiness, achievement, gender, and pedagogical novelty on these variables during this period?

As put by Ganley and Lubienski (2016), we believe that:

Understanding how these constructs are related over time can help educators know if they should specifically target student's confidence or interest, if it will naturally follow if we raise student's achievement, or if we need to address confidence, interest and achievement because they are mutually reinforcing (p. 182)

Since the participants in this study were from the same province and region as our previous transversal study on the decline (Hasni and Potvin 2015a; Potvin and Hasni 2014b),we felt that this longitudinal study might be able to show (again) that perceptual constructs should decline during the studied period (first hypothesis [H1]).We also hypothesized, as noted by Bong et al. (2015), that interest would be the most stable of our endogenous construct over time [H2] while self-concept would be as much [H3]. Indeed, since students have more than one S&T teacher during the period in question and that self-concept usually depends on the type and intensity of personal support that these important models provide, we can expect this construct to be more volatile.

(Potvin et al., 2018)
In light of our discussion above, we used the pairing of pre-university science students with elementary school teachers as a structuring principle for an intervention to improve attitude. In the intervention, the students were matched with the teachers over one school year. Two research questions arise from this approach:

Q1: How does attitude change among teachers who received continuous support through pairing?

Q2: How does intention to teach science topics (covered during the intervention) change?

(Marec et al., 2021)

Participants

Participants in this study are elementary and secondary school pupils aged 10–14 years from diverse French-speaking schools of the greater Montréal area in Canada. The schools were located in different socio-economic environments from well-privileged to under-privileged environments. Pupils were attributed a score of socio-economic status corresponding to the rating of their school from 1 (most privileged) to 10 (less privileged) by the Committee for the Management of Montreal School Tax (2018). Pupils were recruited through the reservation system of the Planétarium Rio Tinto Alcan, where the programming workshop took place. The Planétarium Rio Tinto Alcan is a public institution whose mission is to educate and inspire the general public about astronomy. It is part of Espace pour la vie (Space for Life), Canada's largest natural science museum complex. All teachers who booked the programming workshop were invited to obtain parents' consent and to participate in the study with their class. In total, 188 pupils participated in the study.

(Allaire-Duquette et al., 2022)

Participants

Participants were 182 students (95 boys and 87 girls) who studied in 6 seventh-grade classrooms randomly selected from three junior high schools-two classes from each school. We selected the three schools from a pool of junior high schools in which mathematics was taught in heterogeneous classrooms with no grouping or ability tracking. Each integrated junior high school included students from different socioeconomic status as defined by the Israel Ministry of Education. The classes were similar in size, and students' mean age was 12.4 years.

(Kramarski et al., 2001)

Participants were 81 pre-college Israeli students (N = 54 males and 27 females; mean age = 22.33 (SD = 1.74) ranging from 19 to 27 years old) who studied an advanced course in "mathematical functions."

(Mevarech et Fridkin, 2006)

Sample

The study sample consisted of 517 seventh-grade students—252 males (49%) and 265 females (51%)—from two state schools in Istanbul (the most densely populated city in Turkey).

(Özcan et Eren Gümüş, 2019)

2.1 | Participants

We recruited 123 healthy participants from public schools: 64 children (27 males, M ± SD = 9.8 ± 0.56 years, range = 9–10 years) and 59 adolescents (20 males, M = 16.4 ± 0.52 years, range = 15–17 years).

(Delalande et al., 2019)

Participants

Twenty fourth-grade children (14 girls, six boys; average age = 10 years 3 months, SD = 5.45 months; 18 right-handed, two left-handed) from an elementary school in Caen, France, and 20 young adults (16 women, four men; average age = 20 years 6 months, SD ± 1 year 6 months; 19 right-handed, one left-handed) from Université Paris Descartes, Paris, France, participated in this study. The proportion of males to females did not differ between the 10-year-old children and the young adults, c2(1) = 0.53, p = .47.

(Borst et al., 2013)

Participants

All participants surveyed were students in the second year of a Degree in Primary Education who were studying the subject of 'Curriculum Development of Experimental Sciences' in the Faculty of Education and Social Work of the University of Valladolid in the academic year 2016/2017.

(Reinoso Tapia et al., 2019)

2.1. Participants

The participants were 488 emerging adults (age range: 17–24 years, M = 21.9) recruited through the Amazon Mechanical Turk participant pool.1 Recruiting participants through MTurk allowed us to test whether the mindset intervention is also effective outside of school contexts, as we would expect if mindset effects generalize in the assumed ways. The mean annual household income for participants was approximately $37,500 and ranged from $10,000 to $80,000.

(Burgoyne et al., 2018)

Participants

Fifty-six children (mean age = 9.2, SD = 0.6, range = 8.1–10.1 years, 43% female), 56 young adults (mean age = 22.4, SD = 2.2, range = 18.0–26.3 years, 51% female), and 56 older adults (mean age = 68.7, SD = 3.0, range = 62.3–76.8 years, 59% female) participated in this study. They were recruited from the subject pool at Saarland University, tested individually by one of the eight experimenters and were paid 60 Euros (~95 USD) for participating in the eight sessions of the study.

(Karbach et Kray, 2009)

These students either failed or received a low score on the Israeli matriculation exam in mathematics (M = 64.62; SD = 18.22) and therefore could not be accepted to university studies, unless they improve their scores on the examination. Thus, the pre-college courses provide a second opportunity for those who failed in high-school.

(Mevarech et Fridkin, 2006)

Participants were selected with the goal of increasing the proportion of lower SES participants in our sample, using income as an approximate indicator of SES.

(Burgoyne et al., 2018)

Participants

The study took place at a large public middle school in Tampa, Florida, during the 2013–2014 school year. Three mathematics teachers and nine of their seventh-grade classes participated. Each of the teachers had taught middle school mathematics for more than 5 years. The participating classes included only students who had received a passing score (3 or higher on a scale from 1 to 5) on the Grade 6 mathematics section of the 2013 Florida Comprehensive Assessment Test (FCAT, Version 2.0; Florida Department of Education, Bureau of K–12 Assessment, 2014), which they had taken in April 2013, near the end of the previous school year. A passing score on this test was achieved by 58% of the sixth-grade students at the participating school and by 52% of the sixth grade students in the state (Florida Department of Education, 2013).

(Rohrer et al., 2015)

The current investigation included 142 participants.

(Agogué et al., 2014)

GROUP ASSIGNMENT

6.1 | Randomization design and group assignment

The experiment utilized a double-blinded 2-level stratified randomization design to ensure equivalent treatment and control groups (Bruhn & McKenzie, 2009). Because the study involved students from different grades, teachers, courses, and schools that may be subject to unique experiences within each class section (e.g., socioeconomic factors, learning experience), the randomization was conducted using each class as an individual stratum for treatment assignment whereby students in the same section were assigned to both the treatment and control conditions (Athey & Imbens, 2017). By eliminating potential sources of differences between course sections, stratified assignment increases our ability to detect smaller treatment differences than with other methods of randomization (Box, Hunter, & Hunter, 2005).8 Student demographics, interest in science, experience with Web media and attitudes towards Western medicine were used to randomize students at the individual level within course strata.

We randomized all potential respondents to a condition but omitted those who did not consent or were absent for the study period, yielding 547 participants in the control group and 534 participants in the treatment group (i.e., N = 1081). Randomization was checked for balance against these factors using an omnibus balance test (Hansen & Bowers, 2008), in which a regression was run with the treatment as the dependent variable on baseline characteristics, and we implemented a joint F-test to establish that all coefficients on baseline variables were jointly equal to zero. We found no significant differences in the baseline characteristics between treatment and control participants.9 In summary, randomization produced balanced groups, and subsequent treatment effects represent the causal effect of the intervention on dependent measures.

(Tseng et al., 2021)

Design

We manipulated practice schedule (interleaved or blocked) and test delay (1 or 30 days). Test delay was manipulated by randomly assigning each student within each class to either the 1-day or 30-day delay (n = 63 for each group). This meant that each class included students at both test delays.

Practice schedule was a counterbalanced within-subject variable. Students in Group 1 received interleaved practice of graph problems and blocked practice of slope problems, and Group 2 received the reverse. Group 1 (n = 59) included four classes (two taught by Teacher A, and two taught by Teacher B). Group 2 (n = 67) included five classes (two by A, two by B, and one by C). Classes were assigned to groups as follows. Two of the classes were designated as "honors/gifted" by the school, and these two classes were randomly assigned to different groups. The remaining seven classes were deemed by the school as being at the same level, and each of these classes was randomly assigned to one of the two groups with the constraint that teachers with more than one participating class had an equal number of classes in each group. The two groups scored similarly well on a test consisting of six multiple-choice problems from the Grade-8 mathematics portion of the National Assessment of Educational Progress, or NAEP (National Center for Education Statistics, 2013), 76% (SD = 22%) vs. 79% (SD = 22%), t(108) = 0.62, p = .54, Cohen's d = 0.12.

(Rohrer et al., 2015)

Levels of mathematics achievement were assessed prior to the beginning of this study.

Each of the six teachers who participated in this study taught one classroom. All the teachers were women who had a similar level of education (B.Ed. major in mathematics), had more than 5 years of experience in teaching mathematics, and had taught in heterogeneous classrooms. The teachers were exposed to a 1-day in-service training as described in the Treatment section. We assigned schools randomly to one of the following conditions:

1. MMT Students studied both mathematics and English with the IMPROVE method (n = 60).

2. UMT Students studied only mathematics with the IMPROVE method (n = 60).

3. Control group: Students did not study with the IMPROVE method; that is, students were not directly exposed to metacognitive training (n = 62).

(Kramarski et al., 2001)

2.1. Research model

This study was based on a quasi-experimental design, which is a quantitative research method. A quasi-experimental design is adopted in cases in which experimental and control groups are not formed randomly; instead, they are formed with already-existing classes (Fraenkel & Wallen, 2000; McMillan & Schumacher, 2010). In this design, the experimental and control groups are compared with each other based on a pre-test to determine whether the groups have equal levels of knowledge and achievement. If the knowledge and achievement levels are equal in both groups, one of the classes is chosen to be the experimental group that follows an intervention program.

In this study, the students in the experimental and control groups were selected from intact 7th grade classes at two different schools. This was necessary because of the limited number of classes (one for each) in each grade in the district where this study was conducted. Students' average grade in their previous science class was taken as their pre-test scores, which were compared using an independent samples t-test. This analysis showed that there was no significant difference in the mean values for the control group (M = 68.00, SD = 14.73) and the experimental group (M = 68.38, SD =15.48), t(98) = 0.126, p =.900.

Both groups took courses on the "Solar System and Beyond", which is a unit in the science curriculum. While the experimental group completed this unit with AR technology, the control group completed this same unit with traditional methods and textbooks. In order to evaluate the experimental and control groups' achievement and attitudes towards the course, the "Science Course Achievement Test" and the "Attitude towards Science Course Scale" were used. In addition, the "Attitude towards AR Activities Scale" was used to determine the experimental group's attitude towards AR applications. All data collection tools were applied at the end of the implementation period. The research model is shown in Fig. 1.

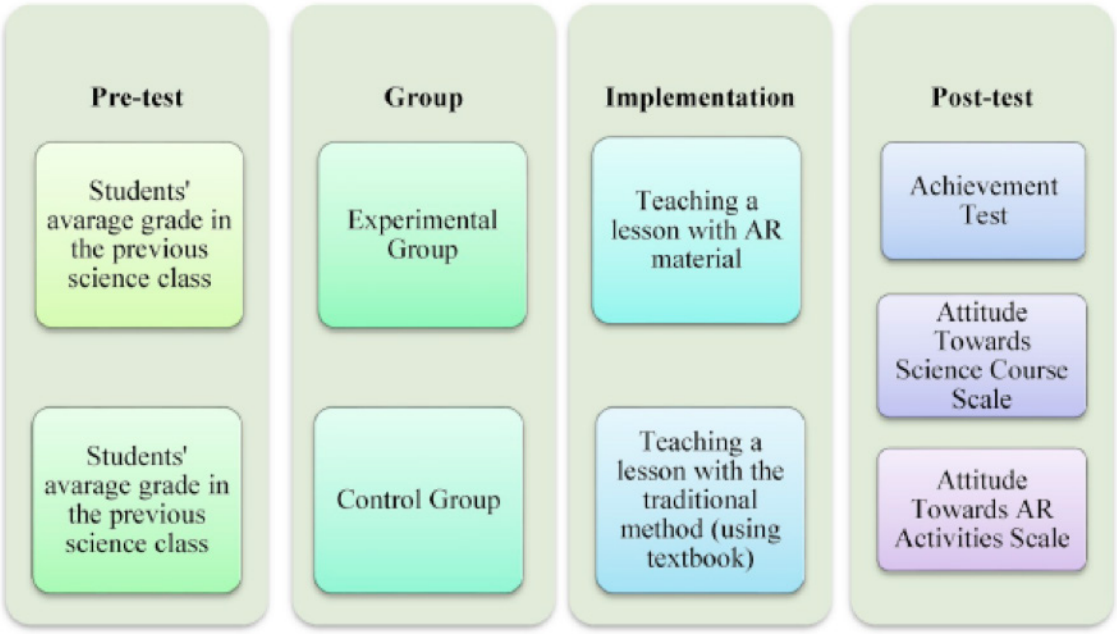



Fig. 1. Research model of the study.

(Sahin et Yilmaz, 2020)

Procedure

Participants were tested class by class (groups of 29 students maximum) under laboratory conditions in the LabUQAM, which is normally used as a teacher training facility. Upon arrival, participants were invited to freely choose a computer station (among the 29 unidentified stations that were available). Verbal instructions were then given to each group as a whole, lasting approximately 6 minutes. Instructions were rendered consistent through the use of a PowerPoint1 (Microsoft Corporation, Redmond, Washington, USA) presentation (15 slides) that presented the goals, explained the procedure, and encouraged participants to "do their best." These instructions were also given verbally by the same presenter each time. After a brief question period, participants put on their individual headphones and began the task (see Figure S3 [available in the supplementary material]: a participant is conducting the task in the laboratory setting). Every student had to take the pretest (step 1), watch the first video (step 2), watch the second video (step 3), and then take the pretest again (retest, step 4). Tasks and videos were imbedded in an E-Prime1 procedure that ensured the continuity of the sequence, contained instructional slides, and provided automatic recordings of participants' interactions with the task, including reaction times.

Participants were thus assigned to one of three possible conditions, depending on the computer station they randomly chose when they arrived. These conditions differed by the choice or order of the two videos. The entire process lasted approximately 45 minutes. Table 1 provides a description of each condition and chronological step, as well as the number of participants for each condition. Note that more subjects were assigned to conditions No. 1 and No. 2 because we initially suspected that significant differences between these two conditions would be slightly more difficult to record.

Table 1
*Description of conditions*

| Condition | Name of the Condition | N | Step 1 | Step 2 | Step 3 | Step 4 |
|---|---|---|---|---|---|---|
| No.1 | *Classical* model | 190 | Pretest | CC | TT | Retest |
| No.2 | *Prevalence* model | 213 | Pretest | TT | CC | Retest |
| No.3 | *Repetition* of traditional teaching | 155 | Pretest | TT | TT | Retest |

Since there are more than two groups to compare, we will use, when applicable, the ANOVA statistical test, which reduces the number of performed tests (instead of using multiple t-tests) and therefore is stricter (less probability of finding a significant result by chance). The ANOVA compares means between groups, so it can resist differences of sizes between these groups. Our experiment also respects ANOVA's assumption of random sample. For instances of comparisons between two groups, we will use t-tests. Effect sizes (partial eta-squared [h2p ] for ANOVAs and Cohen's d for t-tests) are included for all tests and discussed.

(Potvin et al., 2015b)

Procedure

Prior to the beginning of the study, all students were administered the mathematical examination (pretest) followed by the two meta-cognitive questionnaires: the general and the domain-specific. In addition, students fulfilled a short information questionnaire regarding their age, gender, and mathematical score on the matriculation examination. Initial comparisons of the experimental and control groups indicate no significant differences between the groups on age (M = 22.44 and 22.21 SD = 1.623 and 1.833, for IMPROVE and control groups, respectively; F < 1.00; p = .549), mathematics matriculation scores (M = 63.42 and 65.67; SD = 20.78 and 15.80, for IMPROVE and control, respectively; F < 1.00; p = .582), and gender (30 and 24 boys in IMPROVE and control groups, respectively; chi square = 3.87, critical value = 3.84; p = .05).

After the pre-testing, students started to study the solution of maximum and minimum mathematics problems according to the condition to which they were assigned: IMPROVE or traditional instruction. As indicated, the duration of the study was one month.

At the end of the study, all students were administered the posttest and the two meta-cognitive questionnaires: the general and the domain.

(Meravech et Fridkin, 2006)

In the first part of the data collection, students completed the measures of mathematics anxiety, mathematics self-efficacy, and motivation. Students then completed one mathematical problem-solving task, followed by the metacognition experience scale after they solved a problem. An additional three problem-solving questions were then presented in order to acquire a problemsolving performance score. The participants completed the data collection instruments across two class periods (80 minutes) without a break. Mathematical problem-solving performance was evaluated separately by two researchers using a holistic scoring rubric, as explained previously.

(Özcan et Eren Gümüş, 2019)

Procedure

The study was based on a quasi-experimental pretest/posttest design.

Figure 3 shows the data collection process.


**Figure 3.** Data collection process.

At the beginning of the school year and before the intervention, all participants (N = 110) completed the DAS questionnaire (Van Aalderen-smeets & Walma van der Molen, 2013). From the results of the questionnaire, a typological analysis identified four categories of attitude (described above) related to the probability that the teacher would teach S&T, namely 1) high potential, 2) promising, 3) indifferent, and 4) reluctant. Among the teachers in the intervention who volunteered to be interviewed, four to nine representative teachers (i.e., within one standard deviation of the mean) in each category were invited to participate in a semi-directed individual interview at pre-intervention, for a total of 25 teachers. After the intervention, all teachers answered the questionnaire again, and 15 of the 25 teachers participated in a post-intervention interview, depending on whether they remained in their category or migrated from one category to another. The questionnaires and interviews were conducted at the teachers' workplaces.

(Marec et al., 2021)

2.3. Procedure

The study took place in the students' classroom, once a week over an 8 week period. All measures and the intervention were presented via a computer, with each student working on his or her own computer. On the first week, demographic questionnaires and the computerized DCCS and Flanker tasks were administered. The Alien Game was then introduced to the students, who created accounts in the experimenters' online testing system with usernames that the students chose themselves. They then played through the first, tutorial level to learn the basics of the game. For the next 6 weeks, students were asked to sign in and play for 20 min per week. Each week, the next two levels of the game were unlocked. Students' performance in the Alien Game was captured via a log file. Variables in the log file included: time played, levels completed, reaction time (i.e., the time between when an alien first appeared and when it was given food or drink), as well as aliens saved and unsuccessful attempts for each level and overall. At the end of each week, a "Leaderboard" email was sent to all of the students indicating the total number of aliens saved and total play time for all students combined, as well as the usernames of the students who had played the most minutes, saved the most aliens, and had the quickest average reaction times. On the final week, the DCCS and Flanker tasks were administered again, and students were given a posttest questionnaire asking about their experiences with the game.

(Homer et al., 2018)

Procedure

The study consisted of 10 practice assignments, a review session, and a test. Each practice assignment consisted of 12 problems presented on two sides of a single sheet of paper. The 10 assignments included 12 graph problems and 12 slope problems, and the remaining problems were drawn from unrelated topics (fractions, proportions, percentages, statistics, and probability). Teachers presented a tutorial on the graph problems immediately before giving Assignment 1, which included the first four graph problems, and they presented a tutorial on the slope problems immediately before giving Assignment 2, which included the first four slope problems. However, the scheduling of the remaining eight graph and eight slope problems varied. With blocked practice, students saw the remaining eight problems immediately, which is to say that all 12 problems of a particular kind (graph or slope) appeared in the same assignment. With interleaved practice, the remaining eight problems were distributed across subsequent assignments (Figure 3 and Appendix A).

Students received the 10 assignments on Days 1, 6, 14, 32–33, 33 or 35, 35 or 38, 45–46, 72–75, 81–82, and 86–88. Students were asked to complete each assignment before the following school day, and the final practice assignment was collected by teachers on Days 87––89. On the due date for each assignment, teachers presented the solution to every problem with the aid of a slide show created by the authors. As teachers presented the solutions, students were asked to correct their errors. Teachers then collected the assignments. Within three school days, one or more authors visited the school and scored each student's assignment without making any marks on the assignments. The assignments were then returned to the teachers.

Students' scores on the practice assignments do not provide a valid measure of learning because students corrected their solutions before giving their assignments to their teachers. Even if teachers had collected the assignments at the beginning of class, the students might have received help from their parents or other students. This ambiguity is typical of students' mathematics assignments, and many teachers encourage students to seek help with practice assignments.

Yet the scoring of the practice assignments provided an objective measure of the fidelity of the intervention (which consisted solely of the assignments). Most important, these scoring visits to the school revealed that each teacher distributed each of the 10 assignments to their students, and students' self-corrected solutions further demonstrated that the teachers presented the solutions to the practice assignments. (Of course, this perfect rate of teacher compliance might have been achieved because we collected the assignments.) The scoring of the assignments also provided a rough measure of student compliance. When the graph or slope problems were blocked, students averaged 81% correct. In the interleaved practice condition, students averaged 84% correct, and they averaged 82% correct for the last eight of the 12 problems (which were the only problems that were part of an interleaved assignment, as shown in Figure 3). Thus, by this measure, the intervention and the counterfactual produced nearly equal rates of student compliance.

(Rohrer et al., 2015)

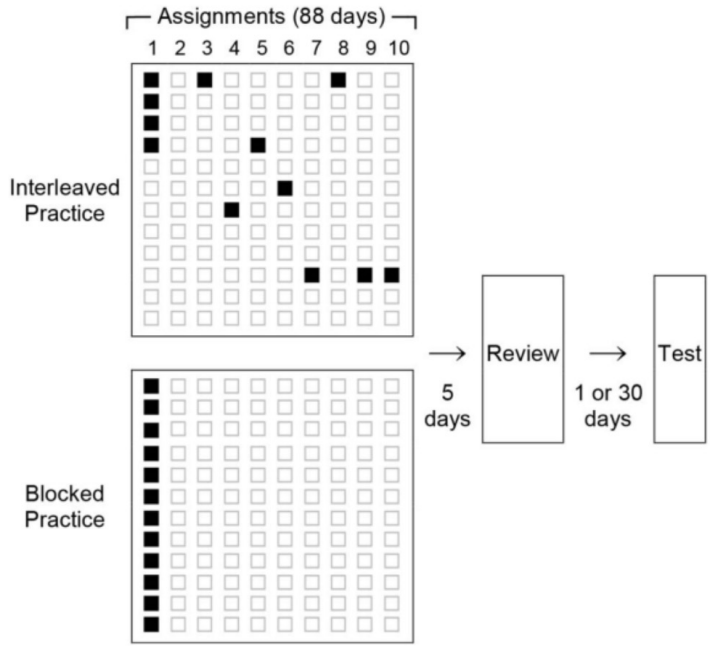



*Figure 3.* Procedure: The 10 assignments included 12 graph problems and 12 slope problems. The 12 problems of each kind were either grouped into a single assignment (blocked practice) or distributed across multiple assignments (interleaved practice). The dark squares indicate the location of the graph problems. The location of the slope problems is given in Appendix A.

(Rohrer et al., 2015)

Protocol

The same questionnaire was administered four times (T1 through T4) during the first 2 years with the same cohort: December 2014, May 2015, December 2015, and May 2016, respectively. These moments were chosen to reflect the state of latent constructs [a] following the first few months of secondary school, [b] as far apart as possible (about 6 months apart), [c] in line with when report cards were issued, and [d] at the very end of the period being studied.

Participants filled out the questionnaires in their classrooms with the permission of their teachers. Instructions were read aloud by the experimenters, and the students were given 35 to 40 min to complete the questionnaires. To ensure anonymity, the completed questionnaires were collected by the experimenter and sent directly to the university for analysis.

(Potvin et al., 2018)

Treatments

To better control the conditions in which teaching was administered, two fundamentally different videos were presented to participants. These videos were prepared by our team (that included an elementary teaching specialist) and both had durations of approximately 5 minutes (4 minutes 28 seconds and 5 minutes 40 seconds) so that teaching time would not be a factor. The first video was the traditional teaching video (hereafter designated as "TT"), and the other was the cognitive conflict video (hereafter designated as "CC"). The TT video could be assimilated to an ordinary science lesson and provided crucial information to participants to help them understand why an object sinks or floats. One of this video's main characteristics was that it avoided negation altogether. No conflicting information was provided. It merely gave "positive" information in the hope that the scientific conception could be understood. It worked essentially by means of (ii) intelligibility, by providing clear explanations designed by a team of three science educators, (ii) plausibility, by showing real experiments, and (iii) fruitfulness, by providing multiple demonstrations of the desired thought process. Figure S1 (available in the supplementary material) shows images from the traditional teaching (TT) video.[...]

The CC video was quite different, multiplying the number of opportunities for cognitive conflict if participants believed that "heavier objects sink more." It used negations and warnings (Houd_e et al., 2000) and did not provide explicit keys to accurate answers. It attempted to trigger dissatisfaction, by providing multiple discrepant events. It provided the opportunity to confront misconceptions by showing multiple cases in which weight distracts from the actual behavior of the balls. Short pauses provided participants with the opportunity to make undisclosed predictions, some which were likely to be contradicted. It sometimes overemphasized examples (see rock/ tanker below) in order to increase the likelihood of cognitive conflicts. Figure S2 (available in the supplementary material) shows images from the cognitive conflict (CC) video.

(Potvin et al., 2015b)

Treatment

Under all conditions, each period in the in-service teacher training included three parts: (a) introduction to the whole class (about 10 min); (b) activities practice (about 30 min); and (c) review with the whole class (about 5 min).

The differences between treatments were as follows:

MMT group: In this condition, the teacher implemented the IMPROVE method (Mevarech & Kramarski, 1997) in both mathematics and English (as a foreign language) classes. A discussion was held in the mathematics class on the similarities and differences in learning a foreign language (English) and a mathematical language (Pimm, 1987, 1996). The teacher presented the studying stages of IMPROVE and held a discussion on the importance of using metacognitive questions in both mathematics and English classrooms. In addition, students discussed the common factors of mathematics language and English, including the importance of understanding vocabulary and symbols, using appropriate strategies to read a text or solve a problem, and reflecting on the text or the solution. [...]

UMT group: In this condition, students were exposed to IMPROVE only in mathematics classrooms. The teacher presented the stages of IMPROVE to the students and discussed the importance of using the method in mathematics classrooms. Students studied in heterogeneous small groups and used the same format as that used in the MMT condition, along with those metacognitive questions described in the preceding paragraphs. The metacognitive questions were used by each individual student when it was his or her turn to solve a problem aloud and by the group as a whole during the mathematical discourse. Also, the teacher used the metacognitive questions when she introduced the new topic to the whole class, reviewed the lesson at the end of the class, and provided help in the small groups.

Control group: In this condition, the control group served as a comparison group with no intervention. Therefore, the teachers of this group continued to teach as they usually did, and students were not exposed to metacognitive training. Each class started with a teacher's short introduction of the new concepts to the whole class. Then students practiced the problems individually without using the metacognitive questions. When students faced difficulties, they consulted the students who sat near them. (In Israel, several students sit at one desk, so they are used to asking for peers' help when they face difficulties.) The students discussed the problem/task and provided help to each other until they solved it. They asked for the teacher's help only when they did not solve their problem. At the end of the session, the teacher reviewed the new concept with the whole class.

(Kramarski et al., 2001)

Treatments

All students studied the course mathematical functions. As indicated, the instructional unit on which the study focuses is maximum and minimum problems. The unit was taught for 12 hours a week during one month (about 50 hours). All students studied the same problems and used the same learning materials. One teacher taught all students. The teacher had more than ten years of experience in teaching advanced courses in mathematics.

Students were randomly assigned into one of two groups and groups were randomly assigned to conditions. One group (N = 38) was exposed to IMPROVE and the other (N = 43) to traditional learning instruction.

IMPROVE–Experimental Group

The IMPROVE method was implemented as follows. The teacher first explained the advantages of using the self-addressed meta-cognitive questioning technique. Then, he modeled the use of the Fcomprehension questions_ in solving an optimization problem and students practiced the comprehension questions while solving the maximum and minimum problems. The same procedure was repeated for introducing and practicing the connection questions, strategic question, and reflection questions. Each type of self-addressed meta-cognitive questioning was practiced for 3–6 sessions, all together 20 sessions. The rest of the time, students continued to use the four kinds of metacognitive questioning during the solution of mathematics problems. Students practiced the problems in individualized settings and the teacher provided assistance as needed. At the end of the session, the teacher reviewed the solution of the mathematical problems by modeling the meta-cognitive questioning.

Traditional Instruction-Control Group

The control group was exposed to the traditional method of instruction, in which the teacher introduced the new concepts to the whole class, and then students practiced the problems relating to the new concepts. As in the experimental group, also in the control group, during the practicing, the teacher provided assistance to students as needed.

Mevarech et Fridkin (2006)

Instruments

We used a quantitative approach to answer our research question. We collected data about (a) students' perception of learning and (b) the eleven studied factors. We designed a measure with four Likert-scale items that was validated by a panel of experts to measure students' perception of learning at the end of each outdoor lesson. Two of the items were positively worded, and the other two were negatively worded. We chose to use an even scale ranging from 1 (strongly disagree) to 6 (strongly agree), and to secure the hypothesis of equidistance between the values, no qualifiers were associated with the values 2 through 5. The items, which were written in French on the questionnaire, were (1) "During this outdoor lesson, I experienced useful scientific learning," (2) "During this outdoor lesson, I did not learn much," (3) "I would probably have learned more by staying indoors today," and (4) "I would learn more by going outdoors more often."

This questionnaire was also used to gather data on two of the 11 situational interest factors we were studying. One factor (positive) was used to measure the students' level of preparation: "I was well prepared for this outdoor lesson." Another factor (negative) was used to measure students' opportunity to make choices: "During the outdoor lesson, I did not have the opportunity to make choices."

We asked the teachers to fill out an online questionnaire that collected data on nine of the 11 factors we were studying within 24 hrs of conducting their outdoor lessons. In the first section of the questionnaire, we also asked them about their outdoor teaching experience ("never taught outdoors before the research," "very rarely taught outdoors before the research," "frequently taught outdoors before the research"), where the lesson was positioned within their lesson sequence (first outdoor lesson, second outdoor lesson, etc.), how long the outdoor lesson lasted (in minutes), and whether a lab technician was present (yes/no). We then used a Likert-scale item with the same values as we described previously for the students' situational interest questionnaire to collect data about the weather conditions during the lesson: "The weather conditions were in all respects favourable for achieving the learning objectives of this outdoor lesson." Finally, the teachers were asked to select the options that applied to the outdoor lesson for each of the following four factors: type of activity (listening to scientific explanations, listening to instructions, identifying a scientific problem, making assumptions, experimenting, observing, modelling), outdoor environment (wooded area, schoolyard, park, watercourse, neighbourhood), scientific discipline/ topic (astronomy, biology, chemistry, geology, physics, scientific method), and student grouping (alone, in pairs, teams of three, teams of four, other groupings, entire class). As there could be more than one option for the same outdoor lesson, the teachers also had to select the relative weighting for each choice (0%, 25%, 50%, 75%, 100%).

(Ayotte-Beaudet et Potvin, 2020)

2.5. Data collection tools

In the current study, the "Science Course Achievement Test", the "Attitude Towards Science Course Scale" and the "Attitude towards AR Activities Scale" were used as data collection tools.

2.5.1. Science course achievement test

This achievement test was based on a test developed by Arici, 2013 and was adapted by the researchers. The test consisted of 20 questions with a reliability coefficient of 0.73. The test questions were based on the Solar System and Beyond unit gains in the 7th grade Science and Technology curriculum. In light of these gains, a table of specifications was formed and questions reflecting each of the gains were created. There were 30 multiple choice questions in the final version of the achievement test. During the preparation process, questions from the Ministry of National Education's terminal exams, course books and Ministry-approved test books were used.

The final version of the test was reviewed by two science teachers and an academician whose expertise is in science education. Any problematic items were revised. The reliability of the achievement test was calculated based on the data obtained from the pilot implementation. If the Cronbach's Alpha coefficient is between $0.00 \leq \alpha < 0.40$ interval, the scale is not reliable; if it is between $0.40 \leq \alpha < 0.60$ interval, the scale's reliability is low; and if it is between $0.60 \leq \alpha < 0.80$ interval, the scale is quite good (Doymuş, 2009). As the achievement test's Cronbach's Alpha value is 0.678, it can be accepted as reliable.

2.5.2. Attitude towards Science Course Scale

In the current study, the 20-item "Attitude Towards Science Course Scale" was used in order to determine the attitudes of students towards science course. It is a 5-Point-Likert scale and composed of 20 items. This scale was developed by Oguz, 2002 and its Cronbach Alpha reliability coefficient is 0.85. [...]

2.5.3. Attitude towards AR Activities Scale

In this study, a scale developed by Kucuk, Yilmaz, Baydas, & Goktaş, 2014 was used to measure students' attitudes towards AR activities. This scale consists of 15 items that reflect 3 factors: satisfaction, anxiety and willingness. Satisfaction relates to students' thoughts on whether AR technology is easy to use and useful for their learning. Willingness reflects students' desire to use the technology in future. If students' satisfaction and willingness levels are high, their attitudes towards AR technology will also be positive. Anxiety relates to any doubts about using AR technology that students might have. When anxiety level is high, students' attitudes are negatively affected. This scale was rated on a 5-Point-Likert scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree) and had an internal reliability coefficient 0.83. The lowest score on the scale is 15 and the highest score is 75. However, students' scores were converted to a 100-point scale.

(Sahin et Yilmaz, 2020)

Factor Structure of the Mastery Experiences in Programming Questionnaire

Mastery Experiences in Programming Questionnaire (MEPQ) was tested and validated in its original French version. An exploratory factorial analysis using the principal component method and varimax rotation was carried out to examine the factor structure of the nine items questioning self-efficacy beliefs. Statistical analyses were performed using SPSS version 27.0 (SPSS Inc., Chicago, IL, USA). The number of factors to retain was determined based on construct validity (Scree test) and interpretability criteria. The resulting factors had eigenvalues of 4.98 and 1.43 and the factor loadings exceeded 0.30, showing minimal overlap among factors. The analysis used varimax rotation with Kaiser normalization and yielded two factors. The first factor (Q_1-Q_6) reflects confidence in accomplishing different programming tasks and was named 'Programming Skills (PS)'. The second factor (Q_7-Q_8; Q_10) reflects the ability to persist despite challenges and was named 'Perseverance when Programming (PP)'. The two factors explained 70.33% of questionnaire variance. The subscales' reliability was calculated by Cronbach α, which reached 0.93 for the programming skills sub-scale and 0.66 for the perseverance when programming sub-scale. Both values were considered satisfactory (Nunally, 1978). Although the perseverance when programming sub-scale would have ideally been greater than 0.70, a threshold of 0.60 can be considered satisfactory when a sub-scale included fewer items as is the case here. This limitation should be kept in mind when interpreting the results. Finally, according to the factor analysis, the 'programming Skills' and 'Perseverance when Programming' sub-scales were moderately related (r = 0.37).

(Allaire-Duquette et al., 2022)

Measuring instruments

We used a questionnaire and a semi-directed individual interview. Among the available questionnaires, we used Van Aalderen-smeets & Walma van der Molen's (2013) Dimensions of Attitude to Science (DAS) questionnaire. Unlike questionnaires on a single subcomponent, such as the STEBI (Science Teaching Efficiency Belief Instrument), developed by Riggs and Enochs (1990), and the CBATS (Context Beliefs about Teaching Science) (Lumpe et al., 2000), the DAS measures seven subcomponents of professional and personal attitude within the dimensions of cognitive beliefs, affective states, and perceived control. The DAS thus provides an overview of attitude, which we complemented with questions about the intention to teach the specific topics covered during the intervention. The questionnaire was forward translated into French by the team of researchers responsible for the study. They worked on the translation independently. They then back-translated it into English and compared their respective versions to arrive at a common translation. The translated version of the DAS uses a six-point Likert scale ranging from "completely disagree" (1) to "completely agree" (6), unlike the original questionnaire, which uses a five-point scale, in order to avoid the neutral point (McMillan & Schumacher, 2006). Exploratory factor analysis (EFA) restricted to seven factors was conducted (given the existence of a theoretical model). A Kaiser-Meyer-Olkin (KMO) indicator (0.79) greater than 0.6 and Bartlett's sphericity test ($p < .001$) below 0.05 confirmed the relevance of conducting the EFA (Tabachnick & Fidell, 2001). The seven subcomponents had acceptable Cronbach alphas, varying between 0.76 and 0.98.

A scale for teachers' behavioral intention regarding specific topics (e.g., electricity) was added to supplement the information collected using the DAS instrument. The new behavioral intention scale had five items, for example, "I am eager to teach the subject of electrical circuits." A KMO index (0.69) greater than 0.6 and Bartlett's sphericity test ($p < .001$) below 0.05 confirmed the relevance of conducting an EFA on this scale (Tabachnick & Fidell, 2001). EFA results gave a single component with loading factors all greater than 0.3, the smallest being 0.69. Cronbach's alpha for the global scale was 0.81. Finally, the questionnaire included questions of a factual nature (personal and academic experience in science, training received, gender, and years of third-cycle elementary school teaching experience).

The qualitative component included semi-structured interviews conducted in order to explain the DAS results in more detail. The interviews provided first-hand access to the teachers' experience and more complete and precise explanations than were possible with questionnaire scores alone. The interviews also gave the teachers the opportunity to freely express themselves about their attitude toward teaching science. The interview protocol included questions that referred to mastery of the notional content, affective states regarding teaching S&T, context-dependency, and intention to teach various scientific topics, including those in the intervention model. The interviews were recorded with the agreement of the participating teachers.

(Marec et al., 2021)

Intrinsic Motivation Inventory

The Intrinsic Motivation Inventory (IMI; Center of Self-Determination Theory, n.d.) is a multidimensional questionnaire grounded on SDT and designed to measure the participants' subjective perception of an activity. Although IMI is mainly used in medical and psychiatric research, it has also found application in, e.g. physics education research (Gustafsson, 2005; Kácovský & Snětinová, 2021).

The questionnaire consists of seven dimensions whose brief description adopted from Monteiro et al. (2015) follows. Interest/enjoyment evaluates interest in and pleasure from an activity. Perceived choice assesses the extent to which individuals feel engaged in an activity because they have chosen it. Perceived competence measures how effective individuals feel in performing a given task. Pressure/tension assesses whether participants experience pressure to succeed in a given activity. Effort/importance reflects how people invest their abilities in what they do. Value/usefulness embodies the idea that people internalise and develop more self-regulatory activities if they perceive the experience as valuable and useful to them. Finally, relatedness concerns interpersonal interactions and refers to a person's feelings associated with others.

Although the instrument is called Intrinsic Motivation Inventory, only the dimension interest/enjoyment directly measures the IM, while the others are considered its predictors (perceived competence, perceived choice, and pressure/tension) or provide supplemental information (effort/importance, value/usefulness, and relatedness). It is common that only some of the dimensions and some of their items are, according to researchers' needs, selected when constructing a tool for a particular study.

(Kácovský et al., 2023)

As Bandura prescribed, a self-efficacy scale should assess the perceived capability to produce given attainments (Bandura, 2006). The standard method of measuring academic self-efficacy is to present problem questions that are similar to actual problems students must solve. Students are then asked to estimate their confidence that they can solve each problem correctly (e.g. Bandura and Schunk, 1981). Thus, our instrument is inspired by the Computer Programming Self-Efficacy Scale for Computer Literacy Education (Tsai et al., 2019). We included in our questionnaire nine items related to the mastery experiences. Six items explore programming skills sub-scale (e.g. 'I am able to test a program on a robot to verify that it works as planned.') and three items explore the perseverance sub-scale related to programming challenges (e.g. "I am able to stay focused even when programming a robot is not going as well as I'd like it to."). In addition, one item regarding enjoyment ("I am able to enjoy programming a robot.") was included to assess emotional state of learners. Pupils rated each item on a 7-point Likert format ranging from 'Not at all able' to 'Definitely able'. We also included three items tapping pupils' satisfaction towards group work (e.g. 'I am satisfied with the way my teammates and I cooperated and shared the work during the activity.') that were rated on a 7-point Likert format ranging from 'Don't agree at all' to 'Totally agree'. One item asked pupils to indicate their prior programming experience. Enjoyment, satisfaction in group work, and prior programming experience have the potential to greatly affect the extent to which pupils gain mastery experiences during the workshop; thus, their scores are used to assess equivalence of girls' and boy's sub-samples before conducting gender comparisons. Finally, an item on general satisfaction asking participants to range their experience on a scale of 1 to 7, where 1 represents 'Did not enjoy at all' and 7 represents 'Enjoyed very much', was included in the questionnaire for the museum's administrative purpose. The full questionnaire is available in English (translated) and French (original) versions in Supplementary Material section.

(Allaire-Duquette et al., 2022)

Factor Structure of the Mastery Experiences in Programming Questionnaire

Mastery Experiences in Programming Questionnaire (MEPQ) was tested and validated in its original French version. An exploratory factorial analysis using the principal component method and varimax rotation was carried out to examine the factor structure of the nine items questioning self-efficacy beliefs. Statistical analyses were performed using SPSS version 27.0 (SPSS Inc., Chicago, IL, USA). The number of factors to retain was determined based on construct validity (Scree test) and interpretability criteria. The resulting factors had eigenvalues of 4.98 and 1.43 and the factor loadings exceeded 0.30, showing minimal overlap among factors. The analysis used varimax rotation with Kaiser normalization and yielded two factors. The first factor (Q_1-Q_6) reflects confidence in accomplishing different programming tasks and was named 'Programming Skills (PS)'. The second factor (Q_7-Q_8; Q_10) reflects the ability to persist despite challenges and was named 'Perseverance when Programming (PP)'. The two factors explained 70.33% of questionnaire variance. The subscales' reliability was calculated by Cronbach $\alpha$, which reached 0.93 for the programming skills sub-scale and 0.66 for the perseverance when programming sub-scale. Both values were considered satisfactory (Nunally, 1978). Although the perseverance when programming sub-scale would have ideally been greater than 0.70, a threshold of 0.60 can be considered satisfactory when a sub-scale included fewer items as is the case here. This limitation should be kept in mind when interpreting the results. Finally, according to the factor analysis, the 'programming Skills' and 'Perseverance when Programming' sub-scales were moderately related ($r$ = 0.37).

(Allaire-Duquette et al., 2022)

Procedure

Every participant was presented with the sinking/floating ball task on a personal computer. At the beginning of the session, each of them was presented with instructional slides. Then they were asked to tell for every trial if the left or right ball was the one that "will have the strongest tendency to sink if it were put in a water tank"—"tendency to sink" means here that the object would at the end be closer to the bottom of the imaginary tank in which the objects would be plunged. The 54 different images (18 "intuitive" (including six "very intuitive"), 18 "counter-intuitive" (including six "very counter-intuitive") and 18 "neutral") were presented in a random order four times to each of the students, for a total of 216 stimuli, with each sequence separated by a short pause. There was a fixation (black "+" sign) of 400 ms between each stimulus, and a maximum delay of 5,000 ms was allowed to produce answers. Participants were asked to give answers as quickly as they could, although it was indicated as more important to give correct answers than fast ones. Participants had to answer by pushing keys 1 and 2 (left ball and right ball, respectively) on the keyboard. There were an equal number of right and left correct answers so that usual biases due to the use of particular fingers or hands (Aoki, Francis & Kinoshita, 2003) were compensated. For each presented stimulus, the E-Prime™ software recorded the order of presentation, accuracy and reaction times.

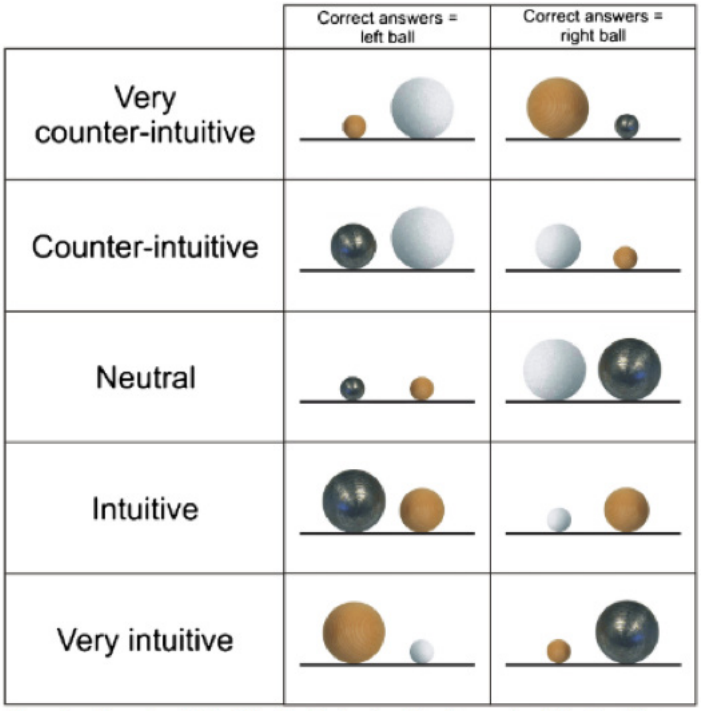



*Figure 1.* Examples of stimuli for the five levels of interference (conditions) and the two possibilities of correct answers (*left*/*right*)

(Potvin et al., 2015a)

Data collection

The data collection was carried out about 10 months after the students had been exposed to instruction involving practical tasks on rock classification. Specifically small groups of students (between two to four per group) were given a collection of five specimens (two metamorphic, two magmatic and one sedimentary), and asked to classify the specimens and explain why they had sorted them as they did. The specimens used were from the authors' private collection and thus not identical to the specimens students may have seen before. This approach was also used by Frøyland et al. (2016) and aimed to encourage students to apply scientific practices versus merely memorising a specimen they had seen before (Stofflett, 1994). The student groups' rock classification performances were video recorded using head-mounted cameras (HD GoPro). Working in small groups ensured that the students' use of observation were more visible through their verbal and physical interactions (Hodson, 1986). After a few minutes, the student groups were asked by the researchers for their conclusions. If necessary, the students were prompted to clarify how or why they classified the specimens as they did. This provided an opportunity to justify conclusions, which is a technique used in investigations of students' understanding of scientific concepts (Kortz & Murray, 2009; Mills, Tomas, & Lewthwaite, 2017; Osborne & Gilbert, 1979). The rock classification task carried out by 19 student groups produced 19 videos, each 5–10 min in length. Below, the video analysis is explained, beginning with the analytical framework.

(Remmen et Foyland, 2020)

Materials

We designed a cognitive task on the PsychoPy™ (v.2020.2.8) software and made it available on the Pavlovia.org online platform. This task could only be run on computers with a keyboard (to record response times uniformly) and took less than 7 min to complete.

After answering simple questions about consent, level of training and/or teaching mechanics, participants read the description of the problems to be solved (See Fig. 2 for a few slides). In a nutshell:

A person simultaneously releases two objects from the roof of a 100-story building, to see which one will hit the ground first. [...] This can happen on Earth (where there is air, like the air we breathe) or on the Moon, where there is no air. [...] You will witness these experiments and try to predict, but never witness, the outcome (the object that appears on your left will hit the ground first [hit the "left arrow" key], the object that appears on your right will hit the ground first ["right arrow" key]; or both will hit the ground simultaneously ["down arrow" key]). [...]

Participants then began. They had to produce 60 answers to each of our randomly ordered stimuli (to avoid sequencing effects), by choosing the "left", "right" or "down" key when the hands opened and revealed the objects. First, for each trial, the hands were closed during three counted seconds, allowing the participant to assess whether the situation was contextualized on Earth or on the Moon (Fig. 4, left). Then, the hands opened (Fig. 4, right) and revealed the two objects. Participants had already been instructed to answer as fast as possible, but not at the expense of providing the best answers they could.

(Potvin et al., 2023)

Measures

We used three measures in this study to assess students' mathematical reasoning and metacognitive knowledge: (a) a pretest that focused on students' mathematical knowledge before the beginning of the study; (b) a posttest that assessed students' mathematical achievement, mathematical explanations, and ability to transfer their knowledge to the solution of a real-life task; and (c) a metacognitive questionnaire.

Pretest of mathematics prior knowledge: To control for possible differences before the beginning of the study, teachers administered a 41-item pretest to all students at the beginning of the school year. The test covered arithmetic knowledge taught prior to the beginning of the study in the following content: whole numbers, fractions, decimals, and percentages. The test was based on multiple-choice items regarding basic factual knowledge and open-ended computation problems.

Scoring: For each item, students received a score of either 1 (correct answer) or 0 (incorrect answer) and a total score ranging from 0 to 4 1. The Kuder-Richardson reliability coefficient was .87.

Mathematics posttest: A 72-item test assessed students' mathematics achievement. The posttest covered the following topics: rational numbers, identification of rational numbers on the number axis, operations with positive and negative numbers, order of operations, and the basic laws of mathematics operations. The test was composed of two kinds of items; problems of both kinds were mixed. One kind (64 items) was based on multiple-choice items regarding basic factual knowledge and open-ended computation problems; the other kind (8 items) was specifically designed to assess students' mathematics reasoning. The 8 items included problems that ask students to draw conclusions about possible outcomes, make algebraic generalizations, evaluate mathematical expressions, and decide which mathematical laws are appropriate for solving them. On those 8 items, students were asked to give a final answer and to explain their reasoning in writing. To gain a deeper understanding of students' mathematics reasoning, we analyzed separately students' mathematics explanations provided for the 8 items.

Scoring for the jnal answer: For each item, students received a score of either 1 (correct answer) or 0 (incorrect answer) and a total score ranging from 0 to 72. The Kuder-Richardson reliability coefficient was .87.

Scoring for mathematical explanations: A scoring procedure was adapted from the holistic scoring rubrics developed by Cai, Lane, and Jakabcsin (1996). For each explanation, students received a score between 0 and 2, and a total score ranging from 0 to 16. For example, "In the following item, 23 . . . (-2)3, write the sign >, < , or = so that a correct statement will be received. Explain your answer." A score of 0 indicates incorrect explanations or explanations that are irrelevant to the task (e.g., incorrect response: "23 = [-2]3 because when there is a minus in brackets in powers, the minus becomes +"). A score of 1 indicates an explanation that has some satisfactory elements but may omit significant parts of the problem (e.g., "23 > [-2]3 because when the exponent of the power is odd, the result will always be negative"- nothing was mentioned about 23 or about the sign >). A score of 2 indicates a correct response with a clear, unambiguous explanation of one's mathematical reasoning (e.g., "23 > [-2]3. when the exponent of the power is odd and the base of the power is positive, the result is positive. When the exponent of the power is odd and the base of the power is negative, the result will be negative even with brackets. A positive number is always bigger than a negative number")

(Kramarski et al., 2001)

Mathematics self-efficacy scale. The mathematics self-efficacy scale was based on the scale used in the PISA2003 student survey. The item stem of "How confident do you feel about having to do the following mathematics tasks?" was followed by six specific types of mathematics activity: calculating the number of square feet of tiles needed to cover a floor; calculating how much cheaper a TV would be after a 30% discount; using a train timetable to work out how long it would take to get from one place to another; understanding graphs presented in newspapers; finding the actual distance between two places on a map with a 1:100 scale; and calculating the fuel mileage of a car. Responses to the items consisted of a four-point Likert-type scale ranging from 1 ("I am not at all confident") to 4 ("I am totally confident"). Validity of the scale was investigated by Lee (2009) by using PISA 2003 data with 41 countries, including Turkey in a factor analysis. In this study, Lee (2009) concluded that the items of this scale constituted an independent construct. Based on the data from the current study, the scale is valid for Turkey as well, the internal consistency coefficient was 0.81 for the whole scale and confirmatory factor analysis (CFA) results indicated that it also had acceptable fit indices (v2/SD=1.20, Goodness of Fit Index (GFI)=0.99, Adjusted Goodness of Fit Index (AGFI)=0.98, RootMean Square Residual (RMR)=0.03, Root Mean Square Error of Approximation (RMSEA)=0.03, and Comparative Fit Index (CFI)=0.99).

(Özcan et Eren Gümüş, 2019)

Mathematics anxiety scale. Five items used in the PISA 2003 student survey were administered in this study to form a mathematics anxiety scale. Five mathematics anxiety items were presented with responses of a four-point scale (strongly agree; agree; disagree; strongly disagree): "I get very nervous doing mathematics problems"; "I get very tense when I have to do mathematics homework"; "I often worry that it will be difficult for me in mathematics classes"; "I feel helpless when doing a mathematics problem"; and "I worry that I will get poor grades in mathematics." The results of Lee's (2009) study showed that the items of this scale constituted an independent construct and this result is valid for Turkey. Based on the data from this study, the internal consistency coefficient was 0.84 for the whole scale and CFA results indicated that it also had acceptable fit indices (v2/SD=1.93, GFI=0.98, AGFI=0.95, RMR=0.03, RMSEA=0.06, and CFI=0.99).

(Özcan et Eren Gümüş, 2019)

Metacognitive experience scale. The metacognitive experience scale developed by Efklides, Kiorpelidou, and Kiosseglou (2006) was used in this study. This scale used in a prospective (before solving a presented problem) and retrospective (after solving a presented problem) manner: the retrospective part was used in this study. As soon as the problem was solved, the following questions were answered by participants: How familiar were you with the problem? How well did you understand what is required by you to do? How difficult did you feel the problem was? How much effort do you think you had to exert in order to solve the problem? How correctly did you think you could solve the problem? Answers were given on a four-point scale: 1=not at all; 2=a little; 3=quite a lot; 4=very. The internal consistency coefficient of these questions was found to be 0.84 within the context of this study and CFA results indicated that it also had acceptable fit indices (v2/SD=1.93, GFI=0.98, AGFI=0.95, RMR=0.03, RMSEA=0.06, and CFI=0.99).

(Özcan et Eren Gümüş, 2019)

Mathematics motivation scale. The mathematics motivation scale used in the PISA 2003 student survey was administered in this study. This scale includes eight items, four each on internal and external motivations. A sample item is "I enjoy reading materials on mathematics" (Association of Educational Research and Development, 2005). Explanatory and CFAs of this scale were conducted; the internal consistency coefficient was 0.89 for intrinsic motivation, 0.87 for extrinsic motivation, and 0.91 for the combined scale. CFA results for the combined scale indicated that it also had acceptable fit indices (v2=2.89, GFI=0.97, AGFI=0.96, RMR=0.05, RMSEA=0.08, and CFI=0.97).

(Özcan et Eren Gümüş, 2019)

Each participant was given ten minutes to generate individually and to write down in silence as many original solutions as possible to the following problem: "Ensure that a hen's egg dropped from a height of 10 m does not break." The experimentation occurred at the beginning of a lecture, with all the participants in the same classroom. The task was administered silently and individually, and participants had to write down their solutions using short sentences.

(Agogué et al., 2014)

2.2.1. Mindset

This 3-item questionnaire assesses whether the participant believes that their intelligence is fixed or malleable (Yeager et al., 2016). Participants respond to items such as "You have a certain amount of intelligence and you really can't do much to change it" using a 6-point Likert scale, with response options ranging from "Strongly disagree" to "Strongly agree." Higher scores on this measure correspond to more growth mindset.

(Burgoyne et al., 2018)

2.2.2. Grit

This 8-item questionnaire assesses trait-level perseverance and passion for long-term goals (Duckworth & Quinn, 2009). Participants respond to items such as "I often set a goal but later choose to pursue a different one" using a 5-point Likert scale, with response options ranging from "Very much like me" to "Not like me at all." Higher scores on this measure correspond to more grit.

(Burgoyne et al., 2018)

2.2.3. Locus of control

This 28-item questionnaire assesses the extent to which the participant believes that their academic performance is a result of internal or external factors (Trice, 1985). Participants report whether items such as "College grades most often reflect the effort you put into classes," or "I have taken a course because it was an easy good grade at least once" are true or false as they relate to themselves. Higher scores on this measure correspond to more internal attributions.

(Burgoyne et al., 2018)

We developed and validated a survey instrument designed to measure participants' epistemic vigilance (i.e., critical awareness) as a theoretical output of the individual-level effects of critiquing scientific claims for potential errors. Specifically, we designed the survey items to measure the student's vigilance when evaluating the claim's source and the claim itself. The questions on the claim's source evaluated students' perceptions of the author's credibility, the media source's reliability, and the perceived expertise of the author (e.g., "does the webpage appear to be a reliable source of information about the topic?"). The items assessing the claims themselves evaluated students' perceptions of the author's reasoning, and the quality of his arguments based on scientific reasoning and usage of scientific evidence (e.g., "would you share this author's claims as good information to know?"). The instrument validation process included multiple iterations of cognitive pre-testing, an initial pilot study and further factor analysis to ensure the validity and reliability of our epistemic vigilance construct (see Supplementary Materials).6 Our final measure was based on the simple mean of nine multiple-choice items that utilized a 5-point Likert rating scale, all of which loaded on a single factor representing epistemic vigilance.

(Tseng et al., 2021)

Materials

Students received graph problems and slope problems, and no student saw the same problem more than once during the experiment. Graph problems required students to graph a linear equation of the form, $y = mx + b$, where m and b were nonzero single-digit integers. Examples include $y = 2x - 1$, $y = -x + 3$, and $y = -3x + 2$. Each problem included the instruction "Graph the equation" and was accompanied by a Cartesian grid. Students were permitted to use any appropriate method, but they were taught to find points by substituting at least two x values into the equation and find the corresponding y values. For example, for the equation $y = 2x - 1$, the substitution of $x = 2$ yields $y = 3$. For slope problems, students found the slope of the line passing through two given points on the Cartesian plane. Each problem began with the instruction, "Find the slope of the line that passes through the points" followed by a pair of points such as "(1, 5) and (8, 9)" or "(3, 1) and (9, −4)." Students were taught to find the slope by calculating Dy/Dx, which is known colloquially as "rise over run." For example, the line passing through points (1, 5) and (8, 9) has slope 4/7. In every slope problem, the two given points had integer coordinates, and the slope equaled a nonzero fraction between −1 and 1.

(Rohrer et al., 2015)

Metacognitive questionnaire: The metacognitive questionnaire, adapted from the study of Montague and Bos ( 1990), assessed students' metacognitive knowledge regarding their specific and general problem-solving strategies. The questionnaire includes 25 items: 6 items refer to strategies used before the solution process (e.g., "Before I solve a problem in mathematics, I try to say it in my own words"); 5 items refer to strategies used during the solution process (eg, "When I solve a problem in mathematics, I organize the data in a table"); 7 items refer to strategies used at the end of the solution process (e.g., "After I solve a problem, I check whether the answer is logical"); and 7 items refer to general problem-solving strategies used during cooperative learning (e.g., "When I talk about the problem with a friend, it's easier for me to understand it").

Scoring: Each item was constructed on a 5-point, Likert-type scale ranging from 1 (never) to 5 (always) and a total score ranging from 25 to 125. Cronbach's alpha reliability coefficient was .83.

(Kramarski et al., 2001)

Domain Specific Meta-Cognitive Knowledge Questionnaire (DSMK-Q)

A 24 item questionnaire adapted from the study of Montague and Bos (1990) assessed students' meta-cognitive knowledge in the area of solving maximum and minimum problems. The questionnaire refers to the use of strategies prior to, during, and after the solution of such problems.

Scoring: Each item was scored on a five-point Likert type scale ranging from never (1) to always (5). Alpha Cronbach equals .85.

(Meravech et Fridkin, 2006)

General Meta-Cognition Questionnaire (GMC-Q)

The meta-cognitive awareness inventory (MAI) developed by Schraw and Dennison (1994) was used to assess students' general meta-cognition. The MAI is a 52 item questionnaire composed of two parts: knowledge of cognition, and regulation of cognition. The first part includes 17 items referring to meta-cognitive knowledge: declarative, procedural, and conditional. According to Schraw and Dennison (1994), "declarative knowledge refers to knowledge about one's skills, intellectual resources, and abilities as a learner. Procedural knowledge refers to knowledge about how to implement learning procedures (e.g., strategies). Conditional knowledge refers to knowledge about when and why to use learning procedures" (p. 474).

The second part of MAI includes 35 items referring to regulation of cognition: planning, information management, monitoring, debugging, and evaluation. The operational definitions of these categories are as follows:

"Planning: planning, goal setting, and allocating resources prior to learning;
Information management: skills and strategy sequences used on-line to process information more efficiently;
Monitoring: assessment of one's learning or strategy use;
Debugging: strategies used to correct comprehension and performance errors;
Evaluation: analysis of performance and strategy effectiveness after a learning episode" (Schraw & Dennison, 1994, pp. 474–475).

Scoring: Each item was scored on a five-point Likert type scale ranging from never (1) to always (5). Alpha Cronbach equals to .87.

Domain Specific Meta-Cognitive Knowledge Questionnaire (DSMK-Q)

A 24 item questionnaire adapted from the study of Montague and Bos (1990) assessed students' meta-cognitive knowledge in the area of solving maximum and minimum problems. The questionnaire refers to the use of strategies prior to, during, and after the solution of such problems.

Scoring: Each item was scored on a five-point Likert type scale ranging from never (1) to always (5). Alpha Cronbach equals .85.

(Meravech et Fridkin, 2006)

Questionnaire design

The questions of the designed questionnaire were based on data obtained from interviews with other teachers, class observations and exams from previous years. Once the difficult topics and concepts were identified, we designed a first version of the questionnaire that included open, semi-open and closed questions. For the validation of the questionnaire, it was first submitted to trial and review by two professors of the department who also teach similar subjects. Subsequently, we provided this first version of the questionnaire to 15 upper-level students, and after its completion, several questions were reviewed. The final version of the questionnaire comprised 12 questions that enquired about conceptual and procedural aspects of the respiratory process, of which 7 were open, 2 semi-open and 3 closed (Appendix 1).

The open questions allowed us to collect qualitative data concerning students' knowledge of external and internal respiration: definitions of breathing, diffusion, cellular respiration, etc.; their ability to distinguish between inspiration and expiration; the understanding of the gas exchange process; their ability to identify the molecules and structures that are involved in respiration, and all the physical processes; and finally, their ability to represent in a schematic way the complete process of respiration. On the other hand, the closed and semi-open questions provided us with quantitative data on the educational stage in which they became familiar with these concepts and the methods or resources used to do so. Additionally, we included a section in the questionnaire dedicated to socio-demographic data, such as age, sex and pre-university itinerary leading up to the Degree in Primary Education, and two more items to mark the research stage number (S1 or S2) and an identification number (e.g. PIN number or personal identity number) that would allow us to correlate the questionnaires completed by the same student in the two stages.

(Reinoso Tapia et al., 2019)

Mathematical Examination

The mathematical examination is constructed of two parts (ten problems). The first part includes five open-ended mathematical problems that examined students' ability to solve maximum and minimum problems. Students were asked to solve the problems and specify all the solution steps. The second part presents five correct mathematical propositions. Students had to provide mathematical justifications for the propositions and explain their reasoning in writing. Examination time was three hours.

Two versions of the examination were constructed: one was used as a pretest, and the other as a posttest.

Scoring: The scores on each problem ranged from 0–10. Thus, the total scores ranged from 0 –100. For the sake of simplicity, all scores, including those of mathematical knowledge and those of mathematical reasoning were transformed into percent correct answers.

Alpha Cronbach reliability scores were .79 and .69, on the pretest and posttest, respectively.

Mevarech et Fridkin (2006)

Mathematics anxiety scale. Five items used in the PISA 2003 student survey were administered in this study to form a mathematics anxiety scale. Five mathematics anxiety items were presented with responses of a four-point scale (strongly agree; agree; disagree; strongly disagree): "I get very nervous doing mathematics problems"; "I get very tense when I have to do mathematics homework"; "I often worry that it will be difficult for me in mathematics classes"; "I feel helpless when doing a mathematics problem"; and "I worry that I will get poor grades in mathematics." The results of Lee's (2009) study showed that the items of this scale constituted an independent construct and this result is valid for Turkey. Based on the data from this study, the internal consistency coefficient was 0.84 for the whole scale and CFA results indicated that it also had acceptable fit indices (v2/SD=1.93, GFI=0.98, AGFI=0.95, RMR=0.03, RMSEA=0.06, and CFI=0.99).

(Özcan et Eren Gümüş, 2019)

B. Item development

One item for each of the 18 statements in the list shown in Table II was written. According to Mattheis and Nakayama [37], several criteria are important to consider when crafting items in a test: the items must represent a wide range of difficulty levels; the items must require an appropriate amount of reading and be at the appropriate reading level for the targeted age group; the items must be devoid of technical terms and scientific jargon that are not understood by children; and the length of the test must allow students to complete it in about 25 min or less. Towns [38] adds that multiple-choice answers should all be about the same length, be similar to each other, be clear of hints or clues that might lead to one of the proposed answers, and be arranged logically or alphabetically, so none of the possible answers stands out. The responses "all of the above" or "none of the above" should also be avoided. Krosnick and Presser [39] encourage the use of simple syntax, and warn against words and expressions with ambiguous meaning. They recommend that a questionnaire should "strive for wording that is specific and concrete (as opposed to general and abstract); make response options exhaustive and mutually exclusive; ask about one thing at a time (avoid double-barreled questions); and avoid questions with single or double negations" (p. 264). As for question order, the authors add that "early questions should be easy and pleasant to answer" (p. 264).

Similarly to most of the instruments listed in Table I, the distractors used as multiple-choice answers for most items in the MPCI-MS were based on a careful examination of past research on students' misconceptions about the lunar phases (see Table S2 in the Supplemental Material [32]). For each item, following recommendations of exhaustivity made by Krosnick and Presser [39], we attempted to propose as many answers as necessary to cover the most common misconceptions found in the literature, or to exhaust all possible answers a child could think of for any given situation described in the items.

(Chastenay et Riopel 2020)

Materials

Three types of items were created: class-inclusion items, subclasses-comparison items, and control items. Each item consisted of one statement and one stimulus. We designed four sets of three stimuli for a total of 12 stimuli (17.8° x 5.1° of visual angle), one set for each of four superordinate classes (square, circle, red, and yellow). In each set, we varied the number of shapes in the two subordinate classes: four versus two, six versus three, or eight versus four. In each stimulus, a horizontal line separated two rows of geometrical shapes, and each row referred to a subordinate class. Two statements were associated with each stimulus: one with "yes" and one with "no" as correct responses (see Figure 1).

In class-inclusion items, stimuli were associated with typical class-inclusion statements, that is, "More A than B" or "More B than A." For example, when eight green squares and four blue squares were displayed, participants could be presented either with "More squares than greens" or "More greens than squares." To respond to such items, the direct comparisons of the subordinate classes' extensions needed to be inhibited. We are confident that the class-inclusion items designed in the present study rely on class-inclusion logic for three primary reasons. First, all of the class-inclusion items follow the three conditions needed to ensure that a problem is a class-inclusion problem (Inhelder & Piaget, 1964): (a) Class A is embedded in Class B, (b) B = A + A__ (in which A__ is not null), and (c) the judgment is of the type "More A than B" or "More B than A." In addition, the types of materials used in the present study are isomorphic to those used in one of Piaget's standardized class-inclusion tasks, that is, round and square tokens of different colors (see Inhelder & Piaget, 1964). Finally, we note that all of the children tested in the present study succeeded on both Piaget's classic class-inclusion problem and Markman's modified class-inclusion problem with two types of materials: (a) blue and red tokens and (b) roses and daisies. All of the children gave correct answers and provided the correct justification for both types of material.

In the subclasses-comparison items, stimuli were preceded by statements requiring direct comparisons of the subordinate classes' extensions. For example, when eight red circles and four yellow circles were displayed, the following statements could be presented: either "More yellows than reds" or "More reds than yellows." Critically, the strategy required to respond to such items was the one that needed to be inhibited to perform the class inclusion items.

For the control items, the same stimuli were used as in the class-inclusion items, but the related statements referred to similarities between the figural properties (i.e., shape or color) of the objects in the two rows. For example, when eight yellow circles and eight red circles were displayed, the following statements could be presented: either "Reds have different shapes" or "Reds have the same shape." In these items, (a) no inhibition was required and (b) the appropriate strategy was different from the one required for the class-inclusion or the subclasses-comparison items.

(Borst et al., 2013)

Materials

The cognitive task was designed with the E-PrimeR software and was inspired by the one designed and used by Shtulman and Valcarcel (2012) and Shtulman and Harrington (2015). It consisted of 40 pairs of textual statements concerning various misconceptions in chemistry (e.g., "The strength of an acid depends on its concentration") whose scientific validity had to be assessed by responding with the index finger of the right hand (if the statement was judged to be true) or the middle finger of the right hand (false). Each pair of statements consisted of one image/stimuli that was congruent with intuition or common sense. This statement could therefore be properly evaluated by most people, whether novices or experts in chemistry. The other statement, on the other hand, was incongruent with intuition or common sense and selecting the correct answer required overcoming a misconception. These statements were presented for a maximum of 10 s using a block protocol (Amaro and Barker, 2006). Each block was composed of four statements of the same condition (congruent or incongruent), followed by a 15 s pause (presentation of a fixation cross). In addition, two mixed blocks composed of congruent and incongruent statements were added to increase the variability of the presentation of stimuli and prevent participants from inferring the structure of the task. These blocks have not, however, been analyzed.

The task was divided into two sessions separated by a short break of approximately 1 min. Each session lasted approximately 5 min during which participants had to respond to 40 stimuli (4 blocks of 4 congruent statements, 4 blocks of 4 incongruent statements and 2 blocks of 4 mixed statements). All stimuli and blocks were presented randomly to avoid presentation bias. However, the first block of the session was always a mixed block in order to add a practice run and allow participants to get used to the task. Figure 3 illustrates a typical example of an incongruent block.

To ensure the reliability and validity of the analysis, stimuli were prepared in pairs and validated by two experienced chemists with a specialization in science education. They were written in such a way as to ensure that they differed only in congruity. Thus, pairs were included in the task only if they respected the following criteria:
1. Relevant to what is taught in chemistry classes;
2. Misconceptions to overcome in incongruent statements were identified as common (or frequent) in at least one science education research article;
3. The statements were scientifically valid;
4. The statements of the same pair dealt with the same scientific concept;
5. Statements from the same pair had the same familiarity and similar complexity of analysis; and
6. The congruent and incongruent statements of a same pair had the same readability and similar lengths [the length criterion was secured by using Flesh-De Landsheere's readibility formula (De Landsheere and Mialaret, 1976)].
7. Also, the task (as a whole) had to contain the same number of true and false statements.

The entire set of statements is available in Appendix.

(Potvin et al., 2020a)

Trials (Fig. 1) started with the presentation of a fixation cross (500 ms), followed by the prime (i.e., a pair of letters, 250 ms) and finally a blank screen that persisted until an answer was provided by the participant within a time limit of 2000 ms. When the participant answered, the fixation cross reappeared (500 ms), then the probe (i.e., a pair of objects, 250 ms) and finally a blank screen that persisted until the participant provided an answer within a time limit of 2000 ms. A visual mask was added between each trial to avoid a transfer of the processes involved from the probe to the subsequent prime (1000 ms). Twelve practice trials, in which participants had to discriminate between pairs of letters and objects that were the same as those used in the experiment, were performed before each of the two experimental blocks (i.e., one block for the lateral condition and one for the vertical condition). Participants performed two blocks of 128 trials in which half of the pairs contained identical stimuli and half contained different stimuli. Two breaks were proposed: one at one-third and one at two-thirds of the blocks. The order of presentation of the trials was pseudo-randomized (with no more than three trials of the same condition or requiring the same answer) and the order of the blocks was counterbalanced.

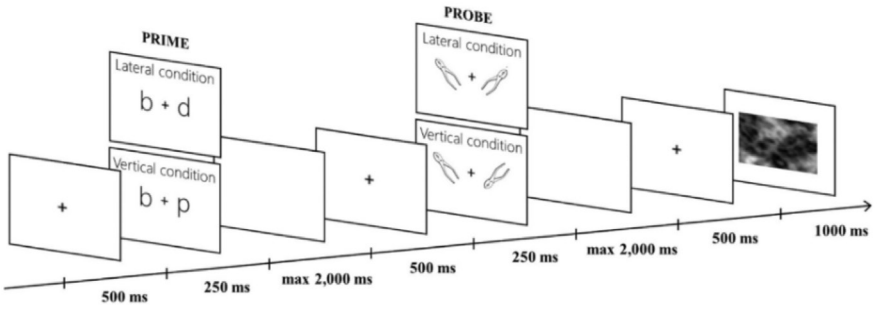



**Fig. 1.** Example of a trial in either the lateral (lower images) or vertical (upper images) condition with two reversible letters serving as primes and two identical objects facing opposite directions serving as probes in Experiment 1.

(Ahr et al., 2017)

Analysis

To verify our hypothesis, the mean reaction time for each participant for each category of stimuli was determined and used to perform paired t tests between the three main categories ("intuitive", "neutral" and "counterintuitive") and between all subcategories (three main categories + "very intuitive" and "very counter-intuitive").

In order to check for negative priming and in accordance with the "prime-probe paradigm" (Babai et al., 2012, p. 766), we divided each of the three main categories into three types: (1) "immediately preceded by intuitive", (2) "immediately preceded by neutral" and (3) "immediately preceded by counter-intuitive". Less data was considered for this last analysis because to be considered, a trial had to be not only answered correctly (probe) but the precedent trial (prime) also had to be answered correctly, which is more restrictive. ANOVAs and t tests were carried out for every combination within the categories.

(Potvin et al., 2015a)

From the initial sample eight pupils were excluded because they did not identify as male or female in the questionnaire. In addition, only data from participants who filled the questionnaire both before and after the workshop was included in the analysis. The attrition rate was 2% (4/188). From the remaining sample of N = 176, outliers were removed based on the three inter-quartiles range rule multipliers. The outliers are believed to be measurement errors, for instance a participant that claims to be fully capable of writing code that can make a robot move but that subsequently claims to not be able to edit, find an error in the code, or predict the output of the code. Four cases were excluded based on that rule, resulting in a 2.2% rate of extreme data (4/176). The final sample included 172 pupils (94 girls and 78 boys).

(Allaire-Duquette et al., 2022)

Data analysis

We used statistical procedures adapted to ordinal variables. We avoided immediately transforming ordinal data into continuous data; we have therefore avoided the priority use of statistics based on averages. Such statistics (ANOVA) will be cited where necessary, as complementary information. Aside from the calculation of frequencies and percentages, we performed three types of data analysis:

1) The questionnaire components and sub-components were statistically validated using a principal component factor analysis, as per techniques used in other research seeking to produce and validate interest or attitude questionnaires (Lamb et al., 2012; Tuan et al., 2005; Wang & Berlin, 2010).

2) Association tests were used to analyze the data according to gender and grade level (Objectives 1 and 2). For gender (binary variable), we used an alternative to the chi-squared test adapted to data tables that vary from the 2 lines x 2 columns format (Fox, 1999), namely the chi-squared likelihood ratio. This test was accompanied by the Cramer V (Øc) measure, which offers an indication of association strength (magnitude). Along with other authors, Howell (1998) notes that "if I had to use only one measure of association, I would choose the Cramer Øc" (p. 182), since this test depends on neither the size of the table that is crossed, nor the size of the sample.

To study the answers to most of the questionnaire items in relation to grade level (all ordinal variables) we used the Goodman-Kruskal Gamma test (g), which also provides a measure of the strength of association (Fox, 1999).

3) To study the relation between the sub-component of General interest in S&T and the other components and sub-components considered in the study (Objective 3), we used correlations and linear regression techniques.

(Hasni et Potvin, 2015)

Since there are more than two groups to compare, we will use, when applicable, the ANOVA statistical test, which reduces the number of performed tests (instead of using multiple t-tests) and therefore is stricter (less probability of finding a significant result by chance). The ANOVA compares means between groups, so it can resist differences of sizes between these groups. Our experiment also respects ANOVA's assumption of random sample. For instances of comparisons between two groups, we will use t-tests. Effect sizes (partial eta-squared [h2p ] for ANOVAs and Cohen's d for t-tests) are included for all tests and discussed.

(Potvin et al., 2015b)

Analysis

In order to provide answers to our research question, we will present, with regard to each hypothesis, the general inter-individual results regarding accuracies (means) for the entire set of data, by competence (schooling, and then relevant teaching experience). For the prevalence of each attractor, we will calculate the percentage of answers that correspond to the conceptual attractor's congruency. For interference, we will calculate the difference in response times between correct answers that are incongruent with conceptual attractors and correct answers that are congruent with it (incongruent > congruent). By doing so, we will obtain a measure of the interference that a conceptual attractor may cause in the production of a correct answer, by comparison with the situation for which the interference cannot occur (congruent). For example, a participant who would correctly answer stimuli 6 and 7 (see Fig. 5; correct answers being "down" and "left" respectively) could experience a bit more trouble answering stimulus 6 because of a possible distracting (interfering) effect of MASS. Thus, the difference in response times between an incongruent (correct) answer to stimulus 6 and a congruent (correct) answer to stimulus 7 could be informative of the interference effect of MASS. Such interference scores cannot be obtained, however, for GALILEO and MASS-DRAG, because there is no correct- and-incongruent answer for those.

Then, for participants of all competency levels, we have performed Pearson's correlations, as well as one-way ANOVAs, with Bonferroni corrections for all posthoc tests in attempts to confirm hypotheses.

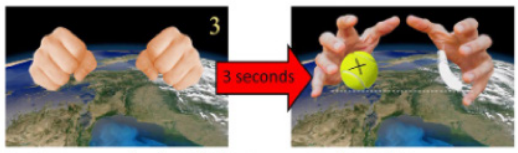



**Fig. 4** An example of a trial (this time on Earth): Trial 40

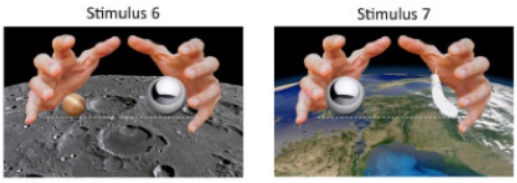



**Fig. 5** Stimuli 6 and 7

(Potvin et al., 2023)

Calculating standardised effect sizes (d) for learning outcomes

With respect to the three learning outcomes, all relevant quantitative data (i.e. M, SD, N) and results of statistical inference tests (i.e. t-values, F-values) were entered onto the same SPSS 23 data sheet. The approach developed by Hedges (1981) and Hedges and Olkin (1985) was used to analyse the data. The standardised effect size computed for every learning outcome in primary studies was d, which quantifies the difference between learning achieved by the group subjected to a serious game and the group subjected to more conventional instruction. To this end, Glass et al.'s (1981) formulas (i, ii, iii, iv, v) revised by Hunter et al. (1982) were used. When means and standard deviations were available, formulas i and ii were used. When a study conducted only a post-test, formula i was used. The mean for the control group (Mctrl) was subtracted from the mean for the experimental group (Mexp), and this difference was divided by the pooled standard deviation for the two groups (SDpooled) obtained with formula v. When a study conducted a pre-test and a post-test, formula ii was used. The post-test mean for the control group was subtracted from the post-test mean for the experimental group, and this difference was divided by the pooled standard deviation for the two groups at post-test to obtain a post-test quotient. The same computation was done for pre-test to obtain a pre-test quotient. Then, the pre-test quotient was subtracted from the post-test quotient to obtain d. When means and standard deviations were not available, either formula iii (t-value) or iv (F-value) was used to compute d.

$$d = (M\,\mathrm{exp} - M\,\mathrm{ctrl})/SD_{pooled} \quad (1)$$

$$d = [(M\,\mathrm{exp} - M\,\mathrm{ctrl})/SD_{\mathrm{pooled}}]_{\mathrm{post-test}} - [(M\,\mathrm{exp} - M\,\mathrm{ctrl})/SD_{\mathrm{pooled}}]_{\mathrm{pre-test}} \quad (2)$$

$$d = t\sqrt{\left(\frac{1}{N\,\mathrm{exp}} + \frac{1}{N\,\mathrm{ctrl}}\right)} \quad (3)$$

$$d = F\sqrt{\left(\frac{1}{N\,\mathrm{exp}} + \frac{1}{N\,\mathrm{ctrl}}\right)} \quad (4)$$

$$SD_{pooled} = \frac{\sqrt{(N\,\mathrm{exp} - 1)SD\,\mathrm{exp}^2 + (N\,\mathrm{ctrl} - 1)SD\ \mathrm{ctrl}^2}}{(N\,\mathrm{exp} + N\,\mathrm{ctrl} - 2)} \quad (5)$$

$$d_{corr} = \left[1 - \left(\frac{3}{4N - 9}\right)\right] * d \quad (6)$$

$$95\%\ \mathrm{CI} = d_{corr} \pm z_{95\%} * \left(\frac{SDd}{\sqrt{N}}\right) \quad (7)$$

(Riopel et al., 2019)

We also conducted within-subjects t-tests, and computed standardized effect sizes using the following formula for Cohen's d:

d = (Mpost-test_Mpre-test)/SDpre-test.

(Burgoyne et al., 2018)

2.6 | Statistical analysis

Relative changes in cognitive performance (i.e. composite IC score) and PFC anatomy (i.e. regional CT and CSA) after training were analyzed using linear models including training group (IC vs. AC), gender (male vs. female) and age group (child vs. adolescent) as categorical factors.

The main effects and interactions were probed with F tests, and post hoc paired comparisons were probed with Tukey's HSD test with Bonferroni correction. A two tailed p value of less than .05 was considered statistically significant.

Relative, instead of absolute, changes were used in the analyses to control for baseline variability. All statistical analyses were carried out with R 2.9 software (http://www.r-proje ct.org/) and 'car', 'effects', and 'Hmisc' libraries.

(Delalande et al., 2019)

Data analysis

Analyses for task switching and the Stroop task were restricted to mean RT for correct responses.3 Practice blocks and the first trial in each block were not analyzed. To control for age differences in baseline performance, we ran ANOVAs based on log-transformed RT (cf. Kray & Lindenberger, 2000). Unless reported otherwise, these results were consistent with those based on mean RT. We also analyzed error rates, but there were no significant interactions with the factor Training on the level of accuracy; therefore, the presentation of results focuses on RT. Data were corrected for multiple comparisons using a Bonferroni correction at $p < .05$. For the remaining tasks (WM, fluid intelligence), the analyses were based on accuracy (% correct) relative to baseline performance at pretest.

To examine the range of transfer effects across training conditions of near and far transfer tasks, we also calculated Cohen's (1977) d, or the standardized mean difference in performance between pretest and posttest (cf. Verhaeghen, Marcoen & Goossens, 1992). That is, the pretest–posttest difference (for each training and age group) was divided by the pooled standard deviation for both test occasions. We then corrected all d-values for small sample bias using the Hedges and Olkin (1985) correction factor (d'). A pretest-posttest effect size d' = 1, for instance, indicates that the mean difference between pretest and posttest corresponds to one standard deviation.

3 For task switching, latencies > 4000 ms were excluded from the analyses (Training: children: 1.43%; young adults: 0.01%; older adults: 0.15%. Pretest and posttest: children: 2.35%; young adults: 0.09%; older adults: 0.81%).

(Karbach et Kray, 2009)

Receptivity of cognitive training was assessed from the relative changes, namely, relative percentage changes = 100 × (post - pre)/ pre, in either the Stroop score (i.e. color-word Stroop task) or the SSRT (i.e. stop-signal task).

For each task, relative changes in either Stroop score or SSRT were screened and cleaned for possible aberrant values using a nonparametric approach: values lower than median - 3 MAD or greater than median + 3 MAD (MAD: median absolute deviation) were considered outliers. Data imputation, based on the median value over the whole sample in each age group, was applied to replace missing data and outliers. Finally, a composite score, combining the relative changes in each task, was then computed to capture a more general IC construct and reduce task-specific variance. This composite score corresponded to the sum of the relative change of each task that was previously cleaned, imputed, and Z-score transformed.

(Delalande et al., 2019)

Results

The data were analyzed as follows. First, one-way ANOVA examines the differences between IMPROVE and control groups on the pretest scores. Then, one-way ANCOVA compares the mean scores on the posttest, controlling for differences on the corresponding pretest. ANCOVA was employed after checking the pre-requisites for running it. All pre-requisites were attained.

Mevarech et Fridkin (2006)

Questionnaire analysis

The questionnaires' answers were evaluated independently by 2 professors of the department, who established a system of coding and categorisation of answers for each of the questions. The coding agreement rate was >90%, and in the case of discrepancies, a third department teacher was required to reach an agreement. Tasks with no or illegible answers were excluded from the analysis. The answers were coded as 'C' for correct and 'I' for incorrect. Among the correct answers, those that fulfilled the criteria of quantity, type or diversity of terms used, were categorised as excellent ('E'), reflecting a greater mastery of the subject.

All the statistical analysis was performed using the statistical package R (version 3.4.4) (R Core Team, 2018). For each item, a descriptive analysis in each of the stages was applied, using as descriptive statistics the percentages of each category together with a confidence interval of 95%. To evaluate the relationship with the demographic variables (age, sex and pre-university itinerary), contingency tables and chi-square contrast were used to evaluate the independence hypothesis in each of the stages. In tables with small expected values, Fisher's exact test was used. To evaluate whether the answers were significantly different in the two moments of time considered, the symmetry hypothesis was contrasted, that is, if the category changes of the analysed variable occurred in both directions with equal probability. For this, the McNemar contrast was used from the contingency table that crosses the two stages. Except for the items of multiple answers, the percentages of improvement, deterioration and non-change were also calculated, along with their confidence intervals.

(Reinoso Tapia et al., 2019)

2.6. Statistical Analysis

We averaged the scores for sports activity in summer and winter and used this value for regression analysis. In addition, we included those with a mean ≤2 (more than one hour a week) into the "more sports activity" group and those with a mean >2.5 (less than one hour a week) into the "less sports activity" group, using an independent samples t-test to evaluate for the hypothesized differences in GMV. We included the following confounds into the analysis known to be an important modulator of GMV: sex (male/female; [28], age [29], BMI [30] and alcohol and cigarettes ([31]. In addition, we corrected for total intracranial volume (TIV) and the quality of images (index of quality rating, IQR) calculated by CAT12 during the preprocessing as covariates. For the design matrix of the t-test design see Figure 1A. For the regression analysis we used an average of the seasonal sports activity scores as the main regressor and the same variables of no interest as for the group comparison. For the design matrix of the linear regression design see Figure 1B.

Since most previous studies investigated effects in an older population (Erickson et al., [12]: n=299, aged on average 78 years; Gow et al., [32]: n=691, aged on average 73 years; Ho et al., [23]: n=226 participants, aged on average 78 years) we tested the interaction age*sports using a 2 by 2 Full Factorial design factorial (sports: less /more sports; age: younger/ older). For that analysis we differentiated the younger and the older group for the t-test by median split. This revealed 475 participants younger than 55.5 years (average: 44.9 years; range: 31-55 years) and 492 participants older than 55.5 years (average: 65.8 years; range: 56-90 years). For descriptive purpose we performed a linear regression for the betas of GMV-analysis in association with age (Figure 2 top) and in association with frequency of sports (Figure 2, bottom).

(Eyme et al., 2019)

They were recruited by the participating teachers and in accordance with the ethics protocol obtained from the university.

(Potvin et al., 2020b)

Ethical approval was granted by the ethics board committees at the authors' university and at other universities included in the study.

(Heilporn et al., 2021)

All of the participants provided written consent or parental written consent and were tested in accordance with national and international norms governing the use of human research participants.

(Borst et al., 2013)

All participants gave written informed consent for participation in the study.

(Jolles et al., 2012)

Ethical Considerations

Information to teachers and parents emphasized that pupils would need less than 10 min to complete the Mastery Experiences in Programming Questionnaire, and that participating in the research was without risk and could even allow pupils to become aware of their own progress in programming and improve their self-efficacy beliefs for programming as a result of the workshop. The pupils were selected for participation in this study via purposeful sampling based on parents' consent to participate. Pupils whose parents did not give their consent were not invited to fill out the questionnaire. However, they were still allowed to participate in the programming workshop at the Planétarium Rio Tinto Alcan without any prejudice. Ethics approval was obtained from the ethics committee (Comite interinstitutionel d'ethique de la recherche avec des etres humains, Université du Québec à Montréal, Canada). Pupils were told that their class was participating in a scientific study aimed at improving programming workshops in science museums. They were invited to complete the Mastery Experiences in Programming Questionnaire before and after the activity if they so wished.

(Allaire-Duquette et al., 2022)

The permission of school principals and teachers was sought and granted prior to data collection; the students participated in the study on a voluntary basis.

(Özcan et Eren Gümüş, 2019)

For instructional domains, significant difference was observed for all domains except the ones related to natural sciences: biology (d = 0.11, 95% CI [0.11–0.33], k = 28) and engineering (d = -0.36, 95% CI [-0.80–0.09], k = 6) for which no significant effect was observed. With respect to methodological moderators, unpublished studies (d = -0.20, 95% CI [-0.83–0.43], k = 3) seemed to yield a lower effect size in favour of games than published studies (d = 0.36, 95% CI [0.24–0.48], k = 67), but this result did not reach significance (p > .05). Random assignment (d = 0.08, 95% CI [-0.13–0.29], k = 35) significantly attenuated (zrandom vs. non-random = 2.75, p = .003) the positive learning effect of serious games compared with lack of randomisation (d = 0.44, 95% CI [0.29–0.60], k = 42). The experimental design had no effect on the magnitude of the effect size (zpost-test only vs. pre-post-test = 0.55, p > .1), with post-test only (d = 0.25, 95% CI [0.07–0.44], k = 27) and pre/post-test (d = 0.32, 95% CI [0.16–0.48], k = 50) studies yielding similar effect sizes.

(Riopel et al., 2019)

The analysis yielded a result that was not significant ($\chi^2(2) = 1.96$, $p > .05$), suggesting that the overall mean effect sizes for declarative knowledge, knowledge retention and procedural knowledge were not different by more than simple sampling error. The three learning outcomes were thus combined for the moderator analyses.

(Riopel et al., 2019)

Table 3. Results from the three-level hierarchical linear model

| Factors | B | SE B | ß | df | t | p |
|---|---|---|---|---|---|---|
| Listening to scientific explanations | .559 | .129 | .209 | 153 | 4.32 | < .001*** |
| Listening to instructions | .204 | .133 | .068 | 153 | 1.53 | .128 |
| Making assumptions | .190 | .204 | .039 | 153 | .93 | .354 |
| Observing | -.229 | .066 | -.161 | 153 | -3.50 | < .001*** |
| Schoolyard | -.052 | .051 | -.049 | 153 | -1.03 | .307 |
| Watercourse | -.105 | .092 | -.049 | 153 | -1.15 | .253 |
| Teacher's outdoor experience | .140 | .087 | .172 | 153 | 1.60 | .111 |
| Presence of a laboratory technician | -.006 | .058 | -.006 | 153 | -.11 | .916 |
| Geology | -.060 | .081 | -.032 | 153 | -.74 | .462 |
| Scientific method | -.081 | .065 | -.058 | 153 | -1.26 | .211 |
| In pairs | .058 | .049 | .051 | 153 | 1.19 | .236 |
| Entire class | .232 | .100 | .128 | 153 | 2.32 | .022* |
| Duration of the outdoor lesson | .001 | .001 | .091 | 153 | 1.75 | .082† |
| Students' opportunity to make choices | .127 | .040 | .152 | 153 | 3.20 | .002** |
| Students' level of preparation | .374 | .044 | .425 | 153 | 8.53 | < .001*** |

*Note.* † $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

(Ayotte-Beaudet et Potvin, 2020)

There was a positive correlation between accuracies and experience ($r = 0.312$, $n = 953$, $p < 0.001$). A one-way ANOVA revealed statistically significant differences in general accuracy between at least two groups ($F(9, 940) = 6.995$, $p < 0.001$).

(Potvin et al., 2023)

One-way ANOVA revealed statistically significant differences in general adherence scores ($F(9, 940) = 14.421$, $p < 0.001$) as well as in interference ($F(9, 819) = 2.306$, $p = 0.015$). The graph thus shows a clear general declining tendency of adherence ($r = -0.32$, $n = 953$, $p < 0.001$), except for university teachers who appear to escape it (however not significantly, possibly because of the rather low number of participants in this group). Post-hoc tests reveal significant differences between most adherence score of lower-than-college students with higher experience groups, however higher groups not being distinguished (from one another) ($p > 0.05$).

(Potvin et al., 2023)

**Table 1.** Mean scores for attitude, the three dimensions of attitude, two subcomponents, and behavioral intention.

| | Intervention group | | | | Control group | | | |
|---|---|---|---|---|---|---|---|---|
| | Pretest | | Posttest | | Pretest | | Posttest | |
| | M | SD | M | SD | M | SD | M | SD |
| Attitude (Q1) | 4.4 | .88 | 4.7 | .61 | 4.6 | .72 | 4.7 | .85 |
| Cognitive beliefs | 4.8 | .72 | 4.8 | .69 | 4.7 | .71 | 4.8 | .58 |
| Affective states | 4.4 | .65 | 4.7 | .82 | 4.5 | 1.0 | 4.7 | .84 |
| **Perceived control** | **3.8** | **.95** | **4.5** | **.81** | **4.3** | **.84** | **4.4** | **.77** |
| **Self-efficacy** | **3.9** | **1.1** | **4.7** | **.91** | **4.2** | **1.0** | **4.5** | **0.9** |
| **Context dependency*** | **3.4** | **1.4** | **2.9** | **1.2** | **2.6** | **1.2** | **2.8** | **1.0** |
| **Behavioral intention (Q2)** | **4.4** | **.95** | **4.9** | **.88** | **4.5** | **.98** | **4.4** | **.98** |

** For this variable, a lower mean signifies less context dependency.*

(Marec et al., 2021)

**Table 3** Standardized path coefficients [β] and *p* values (in parentheses)

| Time | Construct | $R^2$ | Preceding constructs | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Exogenous constructs | | | | Endogenous constructs | | |
| | | | ACH | EASY | NOV | GEN | SC | PUR | INT |
| T1 | SC | .503 | .310 (.000)** | .537 (.000)** | −.019 (.550) | −.101 (.001)** | – | – | – |
| | PUR | .186 | .035 (.392) | .299 (.000)** | .160 (.000)** | −.136 (.001)** | – | – | – |
| | INT | .307 | −.040 (.295) | .380 (.000)** | .247 (.000)** | −.171 (.000)** | – | – | – |
| T2 | SC | .628 | .229 (.000)** | .442 (.000)** | .060 (.036)* | −.034 (.209) | .281 (.000)** | .030 (.362) | .015 (.665) |
| | PUR | .475 | .112 (.001)** | .151 (.000)** | .083 (.016)* | .035 (.280) | −.055 (.164) | .588 (.000)** | −.013 (.760) |
| | INT | .499 | −.045 (.189)** | .242 (.000)** | .269 (.000)** | −.007 (.827) | −.053 (.172) | .082 (.031)* | .417 (.000)** |
| T3 | SC | .660 | .305 (.000)** | .387 (.000)** | −.005 (.844) | −.071 (.007)* | .320 (.000)** | .018 (.557) | −.005 (.875) |
| | PUR | .534 | .069 (.045)* | .139 (.000)** | .017 (.576) | .000 (.994) | .002 (.968) | .653 (.000)** | −.027 (.491) |
| | INT | .552 | .011 (.754) | .266 (.000)** | .267 (.000)** | −.074 (.015)* | −.031 (.409) | .191 (.000)** | .330 (.000)** |
| T4 | SC | .709 | .163 (.000)** | .428 (.000)** | .012 (.635) | −.058 (.015)* | .388 (.000)** | −.018 (.525) | .038 (.234) |
| | PUR | .640 | .020 (.509) | .107 (.002)* | .004 (.885) | .013 (.621) | .004 (.919) | .724 (.000)** | .034 (.341) |
| | INT | .603 | −.010 (.755) | .172 (.000)** | .200 (.000)** | −.026 (.350) | −.025 (.509) | .105 (.001)** | .532 (.000)** |

* $p \leq .05$; ** $p \leq .001$

(Potvin et al., 2018)

Contrary to the pretest, there is no significant effect of gender on PS scores following the workshop after controlling for pretest scores ($F(1, 172) = 0.212$, $p = 0.65$). As was the case before the workshop, no significant difference in post-test scores for PP was found after controlling for pretest scores ($F(1, 172) = 0.041$, $p = 0.84$). Results indicate that after the workshop there were no more differences between girls' and boys' self-efficacy beliefs related to mastery experiences (Table 4).

(Allaire-Duquette et al., 2022)

First, we compare initial (intuitive) versus final (deliberative) responses within the two-response experiment to investigate the causal effect of deliberation within-subject. Consistent with the classical account, we found a significant interaction between headline veracity and response number, $b = 0.36$, 95% CI = [0.2, 0.52], $p < 0.0001$, such that final responses rated false (but not true) news as less accurate relative to initial answers. Moreover, inconsistent with the MS2R account, there was no interaction between political concordance and response number, $b = 0.004$, 95% CI = [-0.16, 0.17], $p = 0.96$, and no three-way interaction between response type, political concordance, and headline veracity, $b = 0.03$, 95% CI = [-0.14, 0.21] $p = 0.72$. Thus, people were more likely to correct their response after deliberation, regardless of whether the item was concordant or discordant with their political beliefs. Naturally, concordance had some effect – people rated politically concordant headlines as more accurate than discordant ones, $b = -0.21$, 95% CI = [-0.34, -0.07], $p = 0.003$ – but this was equally true for initial and final responses.

(Bago et al., 2020)

There was a significant difference in scores for Programming Skills: $t(170) = 2.204$, $p < 0.05$, $d = 1.6$; $\bar{x}$ girls = 3.8 ± 0.2; $\bar{x}$ boys = 4.4 ± 0.2. No significant difference in scores for Perseverance when Programming was found: $t(174) = 0.513$, $p = 0.61$; $\bar{x}$ girls = 5.9 ± 0.1; $\bar{x}$ boys 5.8 ± 0.1). Results (Table 2) show that before the programming workshop girls had lower self-efficacy beliefs regarding programming skills compared to boys, but their perseverance when programming appeared to be similar to that of their male peers.

(Allaire-Duquette et al., 2022)

**Table 1** Comparison of sub-samples of girls and boys on socio-demographic metrics and affective metrics of workshop experience

| Measure | Girls | | Boys | | Stat (df) | Value | *p*-value |
|---|---|---|---|---|---|---|---|
| | x̄ | SD | x̄ | SD | | | |
| Age | 12.0 | .6 | 12.1 | .6 | *t*(167) | 1.490 | $p = .14$ |
| Enjoyment of the workshop | 6.5 | 1.1 | 6.6 | .9 | *t*(169) | .513 | $p = .61$ |
| Underprivileged area score | 5.8 | 2.5 | 5.7 | 2.5 | *t*(170) | .486 | $p = .63$ |
| Satisfaction towards teamwork | 5.7 | 1.4 | 5.7 | 1.4 | *t*(167) | .003 | $p = .99$ |
| First language (# French speakers) | 72/94 | - | 63/78 | - | $X^2(1, 171)$ | .286 | $p = .59$ |
| Prior experience (# novices) | 49/94 | - | 32/78 | - | $X^2(1, 170)$ | 2.533 | $p = .11$ |

*SD*, standard deviation; x̄, mean; *Stat*, statistical test; *df*, degrees of freedom

(Allaire-Duquette et al., 2022)

**Table 1**
Response times (ms) and accuracy rates (%) for the primes and probes in the control and test conditions of the discrimination task in 1st, 3rd, and 5th graders.

| | | RT | | | Accuracy rate | | |
|---|---|---|---|---|---|---|---|
| | | 1st graders | 3rd graders | 5th graders | 1st graders | 3rd graders | 5th graders |
| *Prime* | | | | | | | |
| | Control | 1148 (202) | 1068 (129) | 834 (109) | 86 (12) | 87 (10) | 93 (8) |
| | Test | 1239 (200) | 1175 (142) | 916 (168) | 84 (15) | 81 (14) | 88 (10) |
| *Probe* | | | | | | | |
| | Control | 1195 (190) | 1110 (154) | 893 (117) | 88 (10) | 94 (8) | 90 (9) |
| | Test | 1247 (179) | 1152 (139) | 967 (162) | 81 (11) | 87 (11) | 84 (11) |
| | NP effect | 52 (126) | 42 (81) | 74 (90) | 7 (10) | 7 (10) | 6 (11) |

*Note:* Standard deviations appear in parentheses. The negative priming effect reflects the difference in performance between the two types of probes, which differ only in respect to the type of prime that precedes them. NP, negative priming.

(Ahr et al., 2016)

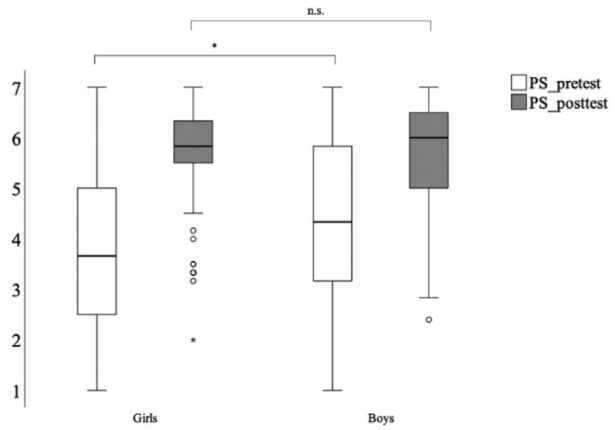
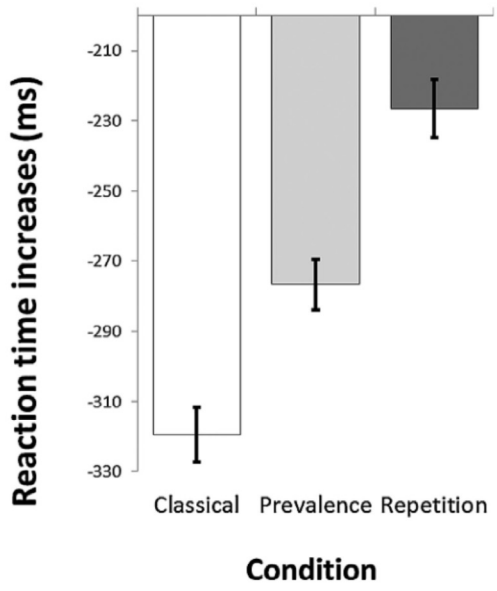



**Fig. 2** Girls' and boys' Programming Skills (PS) score before and after the workshop. *$p < .05$

(Allaire-Duquette et al., 2022)




(Potvin et al., 2015b)

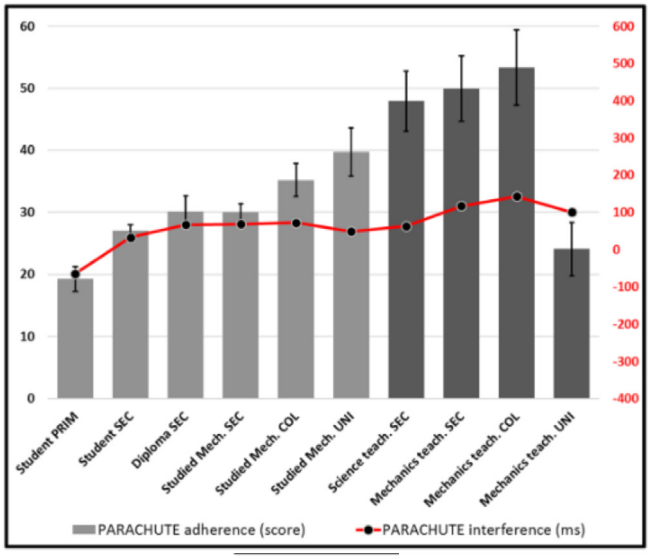



**Fig. 9** Adherence and interference for PARACHUTE

(Potvin et al., 2023)

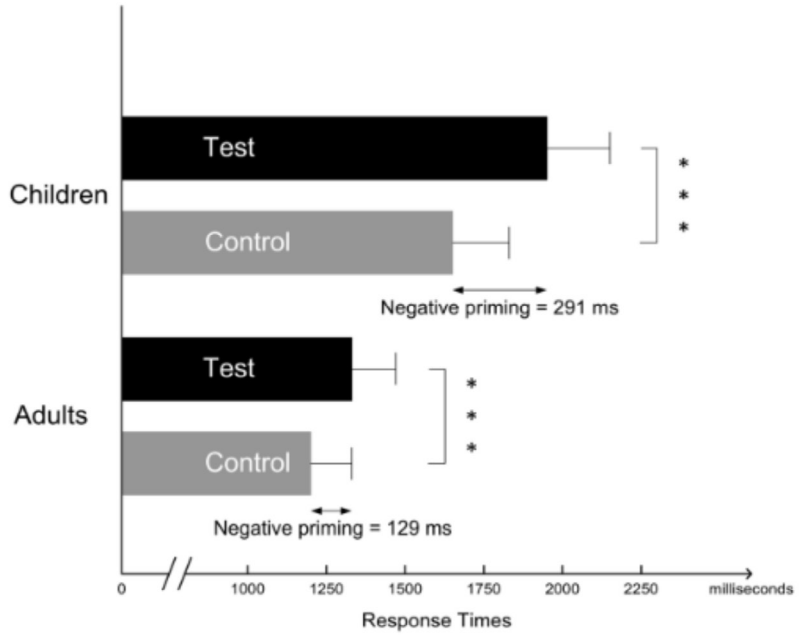



*Figure 3.* Response times on test and control trials in children and adults. Error bars denote 95% confidence intervals. *** $p < .0005$.

(Borst et al., 2013)

We calculated the standardized coefficient with Hox's (2010) formula. The factors that most positively correlated with middle-school students' perception of learning during outdoor science lessons in their schools' immediate surroundings were listening to scientific explanations (ß = .209, p < .001), entire class (ß = .128, p = .022), students' opportunity to make choices (ß = .152, p = .002), and students' level of preparation (ß = .425, p < .001). The results show a significant negative correlation with students' perception of learning when teachers had them observing (ß = -.161, p < .001). Finally, there was a positive correlation with duration of the outdoor lesson (ß = .091, p = .082), but the correlation was not significant (p < .1). The pseudo-R2 (Hox, 2010) was .474.

(Ayotte-Beaudet et Potvin, 2020)

**Table 6** Effect sizes of transitions between school years for each factor

| Level | Elementary | | | Secondary | | | |
|---|---|---|---|---|---|---|---|
| School year | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| *N* | 243 | 375 | 454 | 139 | 572 | 427 | 176 |
| Transition | 5 to 6 | 6 to 7 | 7 to 8 | 8 to 9 | 9 to 10 | 10 to 11 | |
| General interest in school S&T | 0.11* | **−0.53**** | −0.39 | **0.45*** | **−0.33** | −0.23 | |
| Importance of out-of school S&T | −0.08 | **−0.32**** | −0.03 | **0.44** | −0.07 | 0.12 | |
| Utility of out-of-school S&T for society | 0.26 | −0.03 | −0.03 | **0.39*** | 0.14 | −0.23 | |
| Utility of school S&T for everyday life | −0.17 | −0.21 | −0.08 | **0.71** | **−0.61**** | 0.08 | |
| Difficulty of school S&T | −0.04 | 0.20 | **0.69** | −0.14 | 0.05 | 0.13 | |
| School S&T self-concept | 0.21 | **−0.44**** | −0.41 | **0.41** | 0.11 | −0.01 | |
| Attraction to S&T studies and careers | −0.02 | −0.16* | −0.07 | **0.44** | 0.02 | **−0.48**** | |

* $p \leq 0.05$; ** $p \leq 0.005$

(Potvin et Hasni, 2014)

**Table 9** Results of the linear regression analysis for the perceived difficulty of other subjects over school S&T

| | *N* | Intercept | Slope | *p* |
|---|---|---|---|---|
| Arts (not available) | NA | NA | NA | NA |
| English (as second language) | 1,282 | 0.804 | −0.346 | <0.001** |
| French (as first language) | 1,284 | 0.467 | −0.161 | <0.001** |
| Mathematics | 1,208 | −0.012 | 0.004 | 0.883 |
| Physical education | 1,276 | −0.804 | −0.108 | <0.001** |
| Social universe | 1,270 | 0.133 | −0.125 | <0.001** |

* $p \leq 0.05$; ** $p \leq 0.005$

(Potvin et Hasni, 2014)

Q1: change in attitude, dimensions of attitude, and specific subcomponents

To answer our first question, results for the overall concept of attitude did not show significant interaction effect between time and the experimental condition [F(1.97) = 1.1; p = .42; ηp 2= 0.01]. The same was true for the dimensions of cognitive beliefs [F(1.88) = 0.01, p = .99; ηp 2 = 0.00] and affective states [F(1.80) = 2.2, p = .65; ηp 2 = 0.00]. The time x condition interaction effect, however, was statistically significant between the intervention group and the control group with respect to the perceived control dimension [F(1.81) = 7.3; p = .01; ηp 2 = 0.07). We then examined the interaction effect for statistically significant distinctions: analysis indicated that the two groups were different at pretest but not at posttest. Only the intervention group had a significant difference at pretest/posttest (p < .001), with an effect size r of 0.46.

(Marec et al., 2021)

The interaction effect of the factorial ANOVA for behavioral intention was significant [F (1.81) = 5.9; p = .02; ηp 2 = 0.06). The two groups were not different at pretest but were different at posttest. Only the group that underwent the intervention had a significantly different score between pretest and posttest times (p = .01), with an effect size r of 0.30.

A question was asked during the interview about the teachers' ability to "revisit the topic without students" during the following year. In this respect, the vast majority of teachers felt able to do so and wanted to cover the topic next year: "For sure, next year, I'll do it again (1417, posttest)," or "I think I'll be able to do it alone" (1612, posttest).

(Marec et al., 2021)

Table 5 provides reaction times for Pretest and Retest correct answers, and comparing t-test results, for each one of the three conditions. In all treatments, it appears that subjects progressed significantly and substantially (effect sizes considered from small to medium, according to Cohen (1988)), with greater accelerations for classical (d=-0.47), over prevalence (d=-0.42), over repetition (d=-0.36).

An additional ANOVA conducted on pretest reaction times of counter-intuitive stimuli between the three experimental conditions showed that differences were not significant, (F(2, 9394)=0.814, p=0.443, h2p=0.000), supporting the hypothesis according to which students were randomly assigned and were initially equally mislead by the misconception.

Figure 3 presents reductions (negative increases) of reaction times for answers that were accurate on both the pretest and retest. To support his interpretation, one must see that a longer downward bar indicates a greater reduction in response times (i.e. a greater acceleration and therefore better consolidation). A one-way ANOVA was used to test for reaction time reduction differences between the three conditions. These reductions differed significantly across the three conditions, (F(2, 22195)=33.215, p<0.001, h2p=0.003). In post-hoc analysis, all possible differences were significant (p<0.001). Effect size is however very small (Cohen, 1988).

(Potvin et al., 2015b)

Learning outcomes

With regard to hypotheses H1 to H3, findings suggest that serious games are more beneficial, in the context of natural sciences and with equivalent instructional time, than more conventional instructional methods on declarative knowledge gain, knowledge retention and procedural knowledge gain. Thus, hypotheses H1 to H3 are accepted. These findings are in accordance with findings of meta-analyses discussed in the preceding overview (and not specifically related to natural sciences), which have mostly come to the same conclusions. Moreover, it is interesting to note the similarity between mean overall effect sizes computed in the present work and mean overall effect sizes of past meta-analyses. For example, for declarative knowledge gain, an overall effect size of 0.34 was found in the present work, while previous works have found quite similar overall effect sizes of 0.35 for cognitive learning outcomes (Clark et al., 2015), of 0.27 for knowledge learning (Wouters et al., 2013) and of 0.28 for declarative knowledge gain (Sitzmann & Ely, 2011). For knowledge retention, an overall effect size of 0.31 was found in the present work, while previous works have found quite similar overall effect sizes of 0.36 (VanSickle, 1986; Wouters et al., 2013). Sitzmann and Ely (2011) found a slightly lower effect size of 0.22 for this same outcome, a result which could possibly be explained by the fact that Sitzmann and Ely's study focused exclusively on adult trainees and, thus, on a very circumscribed educational context. Although points of comparison are more difficult to establish for procedural knowledge, the overall effect size of 0.41 found in the present work was relatively similar to the overall effect size of 0.37 found by Sitzmann and Ely (2011) metaanalysis. Thus, findings of the present meta-analysis seem to confirm that instruction with serious games, compared with more conventional instruction, is associated with a small to moderate positive overall effect size on science learning achievement. Findings also seem to confirm that the significant effects do not differ when considering declarative knowledge gain, knowledge retention or procedural knowledge. In addition, the impact of serious games on science learning achievement does not appear to be different from their overall impact on learning achievement in other domains of knowledge. This general result, before considering the effects of the moderators, does not support the proposition that natural sciences have a special relation to serious games.

Moderators

With regard to hypothesis H4, findings suggest that serious games are not differentially beneficial on science learning achievement depending on scientific subject area. Thus, hypothesis H4 is rejected. Despite this finding, the discipline of physics appeared to be associated with the highest overall mean effect size, which contradicts some theoretical postulates. Indeed, because physics is a discipline in which misconceptions, or erroneous beliefs about natural phenomena, are particularly well entrenched among learners, it was earlier pointed out that some scholars (e.g. Young et al., 2012) posited that serious games could be detrimental for learning physics because they might induce or consolidate misconceptions. Thus, findings of the present meta-analysis do not support this claim. In addition, the nonsignificant present finding for this moderator could be considered slightly conflictual with previous meta-analyses and meta-syntheses on serious games that examined the moderator effect of subject area. As pointed out in the overview, subject areas compared in these past reviews consisted of various knowledge domains (e.g. mathematics, science, language, etc.) and thus, an inter-domain comparison was conducted. Usually, the conclusion reached was to the effect that learning achieved differed depending on subject area (e.g. Randel et al., 1992; Wouters et al., 2013). This observed nonsignificant difference (while previous meta-analyses observed significant differences for other knowledge domains) could support the claim that all natural science disciplines (physics, chemistry, biology, etc.), because they are based on the same type of quantitative predictive models, have some special relation to serious games that leads them to be homogeneously beneficial.
With regard to hypothesis H5, findings suggest that high school students appear to benefit the most

(Riopel et al., 2019)

Our first research question focused on the change in attitude among teachers who participated in our intervention program based on pairing. In this regard, the results obtained in Table 1 do not show change in general attitude. However, when the three dimensions of attitude are examined separately, interesting observations can be made.

The results of statistical analysis did not show change in the cognitive beliefs section. High scores were observed at pre-intervention for two of the three subcomponents of this dimension, namely, teachers' perceived relevance of teaching S&T (M = 5.2, SD = 0.8) and gender beliefs (M = 2.0, SD = 1.1). Our sample of teachers seemed to believe that science and science teaching were important in students' curricula and did not seem to perceive a difference in gender in terms of ability or potential in science. This result is a reminder of the importance of science education from an early age, which is generally recognized by researchers (Epstein & Miller, 2011; Jarvis & Pell, 2004; Osborne et al., 2003) and now widely shared by teachers.

The results of Table 1 also indicate that no change could be observed for the affective states dimension. For this dimension, the pre-intervention score of the anxiety subcomponent was low (M = 2.3, SD = 1.4) and indicated a certain absence of stress in approaching science.

One of the reasons why our results did not show change in the cognitive beliefs and affective states dimensions, and in general attitude, may be due to a ceiling effect (Roberts et al., 2001).

However, Table 1 indicates that, at the end of the intervention, a significant change was observed with respect to the perceived control dimension, which includes the self-efficacy and context-dependency subcomponents. It is likely that the whole structure of the intervention contributed to the development of the teachers' perceived control. Our results suggest that the participating teachers gained in confidence and self-efficacy. We can assume that the pairing of pre-university students contributed in improving the teachers' self-efficacy. In this regard, 13 of the 15 teachers we interviewed mentioned that the presence of the pre-university students was beneficial to them. The teachers' vicarious experience (one of the four self-efficacy sources) most likely facilitated this new practice: observing pre-university students participating without difficulty in previously conflictual activities (e.g., disorganization surrounding scientific experiments) probably influenced the teachers' ability to be successful in the same activities.
[...]
We acknowledge that our study has some limitations as to its transferability. It was conducted in a particular social context, i.e., all the schools were from disadvantaged communities. In these schools, emphasis is placed on learning basic subjects, while teaching S&T, although deemed relevant by the participating teachers, rarely occupies the time indicated in ministerial directives. In addition, our sample was made up of volunteer thirdcycle elementary school teachers for both the questionnaire and the interviews. In summary, our study showed encouraging results at the end of 1 year of pairing, which allows us to underline the benefits brought to the teachers' practice; however, this study should be continued in order to determine whether the positive effect is maintained the following year and if the intention to teach the notions covered is solidified. From this perspective, it seems to us that studying the relationship between attitude and behavioral intention followed by behavior is a promising avenue of research. It would make it possible to gauge more effectively the development of attitude toward S&T in light of continuing education programs.

(Marec et al., 2021)

General highlights

Not surprisingly, our data support the hypothesis that conceptual understanding of falling body problems improves with age and experience (Fig. 6). Surprisingly, most conceptual gains appear to occur before secondary school, in the absence of any physics training, suggesting that this improvement may be developmental, or that direct experience rapidly corrects certain misconceptions, such as VOLUME. Paradoxically, the improvement slows down when physics training begins. Thus, our hypotheses are only partly confirmed.

A more thorough analysis, including the prevalence (correct answers) and interference (delays or hesitations caused by irrelevant distractors) scores obtained for each conceptual attractor, allows us to go deeper in our interpretation. First, it shows that MASS (Fig. 7) is initially prevalent, as has been noted before (Gunstone & Watts, 1985; Sequeira & Leite, 1991; Whitaker, 1983), but regularly decreases along our competence (age and experience) continuum, suggesting that not only direct interactions with the world, but also training contribute to reducing its perceived cognitive utility and its interference power.

However, it is possible that training efforts may be pushing a bit too hard to promote its rejection, since interference is negative, suggesting that it ultimately contributes to the production of incorrect answers. More than everybody else, high school and college physics teachers appear to be strongly opposed to MASS, except university physics teachers, who also show the best overall scores. It is not unreasonable to suggest here that physics teaching at lower levels may sometimes commit the sin of generalizing an indiscriminate rejection of mass as a relevant variable. To our knowledge, these results are the first to document this shortcoming as convincingly.

[...]

This result is a bit worrying since these teachers are precisely the ones who teach physics to children. Declines that begin at the secondary level suggest that all involved in school business, learners and teachers alike, cultivate this misapprehension of atmospheric falling objects problems. It is somewhat paradoxical, however, that scores on the PARACHUTE attractor (Fig. 9) also generally increase with experience, with physics teachers being the strongest adherents. We can hypothesize that they mostly resort to this misconception only in very special cases, for example to explain why the fall of very oddly shaped objects, like feathers, appears longer in the atmosphere, mostly in cases of falls from low heights. The rather low interference (< 120 ms) by PARACHUTE however suggests that participants do not let it get a lot in their way in most of their reflections.

(Potvin et al., 2023)

The present study explicitly compared interleaved and blocked mathematics practice in a classroom setting and found that interleaved practice produced superior scores on a final test given 1 or 30 days later. Put another way, the mere rearrangement of practice problems improved mathematics learning in the classroom. The study is also the first to demonstrate that the test benefit of interleaving does not diminish over time and perhaps grows larger. Finally, apart from its superiority to blocked practice, interleaved practice provided near immunity against forgetting, as the 30-fold increase in test delay reduced test scores by less than a tenth (from 80% to 74%).

Although the size of the effects might seem surprisingly large for a study conducted in a classroom, the effect sizes observed here are nevertheless much smaller than interleaving effects observed in the laboratory. Whereas the effects in the present study were medium to large (ds _ 0.42 and 0.79), laboratory studies of interleaving have uniformly found larger effects (d = 1.34, d = 1.21, and hp2 = .32; see introduction). In brief, the present findings are an instantiation—not a violation— of the adage that an intervention loses some of its efficacy when it is moved from the laboratory to the classroom (e.g., Hulleman & Cordray, 2009).

Another reason for the large effects of interleaving observed here and elsewhere is that interleaved mathematics practice inherently guarantees that students space their practice. That is, in addition to the juxtaposition of different kinds of problems within an assignment, problems of the same kind are spaced across assignments. However, the review session in the present study meant that even the blocked practice condition provided spacing, although to a lesser degree than that provided by the interleaved practice condition (Figure 3). In brief, the large effect observed here probably reflects the spacing effect, which is an inherent benefit of interleaved mathematics practice, but the contribution of spacing might have been reduced by the use of a review session.

The large effects notwithstanding, the present study has limitations. For instance, although the test problems were novel, the test problems and practice problems had the same format, and the observed effects might have been smaller if the test problems had required a greater degree of transfer. Also, the test benefit of interleaving might have been reduced if the review had included more than one problem of each kind (graph problem and slope problem), simply because a more intensive review session might have benefitted the blocked practice condition more than it did the interleaved practice condition. More broadly, it remains unknown whether the interleaving effects observed here would be found in a study with a wider variety of material and a greater number of teachers and students. Still, the ecological validity of the present study was reasonably good. Students learned from their teachers, the learning phase lasted 3 months, and the 1-month test delay was educationally meaningful.

(Rohrer et al., 2015)

This study provided a rare opportunity for students to confront science information as potentially false claims—a context that students encounter increasingly often in our contemporary information landscape. Furthermore, this study is one of few that experimentally tests the effects of an intervention to support critique and evaluation in a science classroom setting, building upon the extant body of literature on the pivotal role of critique in the assessment of scientific claims. Our results highlight four key points.

First, students' epistemic vigilance was not significantly different between the treatment and control group, but this was mediated by significant attrition in the treatment group. This finding highlighted the need for improvements in the intervention's design and length, as most of the attrition occurred while students completed the intervention activity. Our ad hoc analyses point to several design features of the intervention that may mediate improvements in epistemic vigilance. For example, the lengthy reading guide may have been plagued by slow or unstable Internet connections at school sites. This could be improved by presenting it on shorter consecutive web forms in order to load reliably. Other design issues that may have impacted intervention compliance or contributed to respondent fatigue were the length of the form, placement of visual elements, and usability of the form on various devices and browsers (Jarrett & Gaffney, 2009; Wroblewski, 2008).

Second, when accounting for treatment compliance, there was an increase in epistemic vigilance, which suggests that inducing students to critique the text may be effective in raising critical awareness in the face of misinformation. We found positive effects for students who completed the critical reading intervention, equal to an effect size of 0.10. This is the equivalent of a participant increasing their epistemic vigilance by moving from the 50th to the 60th percentile of the sample distribution. This is a small effect, especially given that the effect was detected using a specialized instrument to measure epistemic vigilance and the outcome was measured close in time to the intervention (Kraft, 2020). However, this intervention would also be considered a "light-touch" intervention—classroom teachers did not receive extensive professional development to conduct the intervention and the intervention itself (i.e., a critical reading scaffold) took less than a single class period to implement. This intervention was low-cost, highly scalable, and transferrable to other subjects (e.g. civics education). Other recent classroom-based interventions in the areas of source evaluation (Bråten et al., 2019) and evaluation of lines of evidence (Lombardi et al., 2018) have achieved larger effect sizes (though these studies are difficult to compare given different dependent variables, measures, and statistical procedures); however, these interventions were intensive, requiring professional development for teachers and multiple class periods for students. In sum, the present work can be considered as proof of concept—when students were prompted to "think slow" with a brief reading guide, they were more vigilant against misinformation if they engaged fully with the activity.

Our study adds to the growing literature that underscores the value of critique in argumentation and specifically, challenging flawed scientific arguments—an area that continues to be neglected in science education. [...] During a time period in which wide-reaching technology can instantly spread truths and falsehoods alike, it is imperative to educate today's youth to be vigilant, before "our knowledge is a receding mirage in an expanding desert of ignorance (Durant, 2011)."

(Tseng et al., 2021)

Discussion

The primary aim of this study was to investigate the usefulness of task-switching training. To answer this question, we examined the amount of near and far transfer of task-switching training in children, young adults, and older adults under different training conditions. Our results identified several important new findings. First, we found evidence for substantial transfer of task-switching training to a structurally similar new switching task after training. Consistent with a prior study (Minear et al., 2002), the reduction of mixing costs from pretest to posttest was much larger after task switching training (mean d' = 1.44) than after single-task training (mean d' = .26). From a theoretical point of view, this finding is particularly important because it shows that the trainability and transferability of executive control processes is not merely mediated by automatization of single-task components (cf. Kramer et al., 1999). In contrast to the Minear et al. study, we also found near transfer of task-switching training on the level of switching costs (mean d' = 1.17).

Second, and particularly interesting from a developmental perspective, the near transfer on the level of mixing costs was most pronounced in children and older adults. Thus, in particular the age groups usually characterized by marked deficits in task-set selection and maintenance were able to transfer training-related benefits to a new task. This finding has important implications for the application of training programs to individuals with executive deficits in the clinical and educational contexts.

Determining the relative training potential regarding different executive control components and the long-term effects of task-switching training is a matter for future research. Also, it may be important to consider individual differences regarding transfer benefits to clarify how the training improves performance, so that it can be optimized when applied to those who need it most (cf. Bissig & Lustig, 2007; van Merriënboer et al., 2006).

(Karbach et Kray, 2009)

Limits

Among the limits of this research is, of course, the small number of participants (N = 17). Indeed, it always remains impossible to conduct studies with complete generalization power, even if the studied population is as small as the number of chemistry professors. Since its size, our sample could therefore be unevenly tainted in ways we did not think about, thus threatening a possible generalizabilition to all chemistry professors. For example, our entire sample is composed of culturally rather uniform (french-canadians) professors. It is also composed exclusively of voluntary participants. And thus, while we are unable to imagine in which direction such characteristics could have skewed our results, we cannot defend the possibility that they did not. Therefore, reproduction through similar yet converging studies would be needed.

Another limit is that, even though our results reached statistical significance, this study did not present results with corrected thresholds, which is increasingly becoming a standard in the field. It is possible that the small number of participants or the reduction of considered pairs of stimuli for analysis (we could only consider pairs when both incongruent and congruent stimuli were answered correctly) explained this restriction, at least in part. However, it is also possible that the very high automatization of scientific knowledge in chemistry professors made it more difficult for us to detect the interference of misconceptions. Indeed, selecting chemistry experts was a bolder methodological choice that made reaching significant thresholds more challenging and, with eventual positive results, more educationally interesting.

(Potvin et al., 2020a)

Limitations and future research

We identify two specific limitations in the current study. First, the study is limited by a lack of long-term measures. While the evidence provided here supports that the program innovations contributed to a short-term increase in fourth grade student interest, selfefficacy, and improved perceptions of who can be a scientist, the study did not support investigating the potential long-term impacts of the program. Future research is needed to determine how programs like Science Is Fun can move from initiating situational interest to cultivating maintained situational interest (Hidi & Renninger, 2006) in the long term. Relatedly, future studies are needed to explore how teachers and demonstration facilitators can work together to integrate informal science demonstrations with ongoing phenomena-based learning experiences that support standards-based learning. Although we have argued for a role for informal science demonstrations in phenomenabased instruction that aligns with current curriculum standards, this study is limited by its methodology in making the case for such a role. Design-based research and case studies may prove useful in exploring such extended collaborations.

A second significant limitation of the experimental design described here is a lack of process measures, primarily the lack of control groups to test the independent effect of each aspect of the Science Is Fun pedagogical innovation (i.e. demonstration-experiments, everyday experiences, near-peer role modeling). We tentatively propose that each component of the Science is Fun demonstrations illicit a specific outcome. Specifically, we hypothesise that the demonstration-experiments address a lack of interactivity while supporting interest; contextualisation of demonstrations within everyday experiences addresses a lack of prior experiences and supports interest, and near-peer role modeling addresses power imbalances while supporting selfefficacy and perceptions of who can be a scientist. Nonetheless, explanations for each model aspect on the motivational variables in the discussion can only be proposed as a possible explanation for the observed positive impact, as our design does not allow for variable control. Future iterations of the research that include surveys or interviews could help reveal fourth-grade audience's perceptions of the near-peer role models and interactive science demonstrations to determine if the theoretical benefits of near-peer role modeling were supported. Likewise, we argued above that self-efficacy can positively influence possible selves, which have been linked to students' and teachers' increased motivation and agency (Harrison, 2018; Markus & Nurius, 1986; Narayanan & Ordynans, 2022). As possible selves are known to be activated by social situations and respond to changes in environment (Markus & Nurius, 1986) and their range expanded by imagined communities (Kanno & Norton, 2003), Science Is Fun's high-school near-peer role models offer potential to promote their fourth-grade students' scientific possible selves. Investigating these relationships may offer further fruitful pathways for future study.

(Howell et al., 2023)

Limitations

Presented results should however be considered with great caution. Even though most were clearly significant, some of them showed small or less than small effect sizes, especially when comparing conditions. This could be expected, considering the nature of the data and some important similarities between our experimental conditions, but is nevertheless a limitation. Therefore, further confirmations must be provided before attempts at implementation. Also, for the moment, our "evidence" was brought about with very young students (grades 5 and 6). We know, for instance, that students are able to resolve certain problems only when their inhibitive capacities are fully developed (Houdé et al., 2011). Thus further confirmation with older subjects would help. Confirmations or disconfirmations with conceptions other than the one tested in our "sink/float" task would also be welcomed. There is no way to confirm with our data that other misconceptions do not follow other logics. It would also be interesting to conduct tests in a more authentic setting, where interactions between learners and teachers can take place. In our idealized design (teaching with videos), some of the important virtues of teaching were absent. We can easily imagine that cognitive conflict would be more meaningful in the form of a dialog.

It is finally important to remind that conceptual change is usually considered as a process that requires quite a lot time to be successful. In this perspective, it is interesting to see that with our rather short pedagogical treatments (less than 45 minutes) we nevertheless were able to produce a difference (on accuracies and on RT), even though the durability of this progress is less than secured. Therefore it is not impossible that we produced learning that some science education specialists might not consider as true conceptual changes. However, we used the same kind of indicators (accuracies, RT) than most studies. But we agree that it would be interesting to see what would be the strength of the effects of our treatments on much longer periods, and in real class settings. We therefore believe that long term research efforts could benefit the field.

(Potvin et al., 2015b)

The present work aimed to determine whether serious games were more effective, compared with more conventional instruction, on science learning achievement. For all three learning outcomes examined (i.e. declarative knowledge, knowledge retention, procedural knowledge), serious games were found to be more beneficial than conventional instructional methods. The effect size of this benefit was found to be small to moderate, which is consistent with previous meta-analytical findings on the effects of serious games in other domains. The present work thus concludes, about the special relationship of serious games to natural sciences, that the overall effect is as significant and with an amplitude comparable with other domains of knowledge.

Moreover, several theoretical and methodological moderators were found to affect the link between instruction with serious games and science learning achieved. Findings of the present work suggest that five moderators' effects were significant (grade level, duration of intervention, level of user control, year of publication and publication status). Among those that were not significant, three moderators showed small consistent variations of mean effect size (subject area, activity level of comparison group, level of realism) that could lead to significance with more studies and larger samples. Furthermore, some findings about moderators are intriguing and require more research and new proposals that could contribute to characterising the special relationship between natural sciences and serious games. Finally, three moderators (ludic content, randomization, experimental design) showed neither significant nor important effect on science learning achievement.

Similar to previous meta-analyses on serious games, the present meta-analysis has limitations. For example, it did not examine some moderators frequently found in the literature to have a significant effect between instruction with serious games and learning achieved, such as participants' gender, grouping during game play, or overall quality of the game. It also did not analyse the effect of serious games on other variables, such as motivation.

(Riopel et al., 2019)

Conclusions

In our research, we started from the self-determination theory and used the 'laboratory version' of Intrinsic Motivation Inventory to assess the intrinsic motivation of upper secondary students during their hands-on practical work in the laboratory at our faculty. Our primary goal was to assess how strong predictors of intrinsic motivation are the latent variables effort/importance, pressure/tension, perceived competence, and value/usefulness. Our results show that the strongest predictor of intrinsic motivation is the subjectively perceived value/usefulness of practical work. In students who label themselves as less diligent in physics and those who do not plan studying STEM at university, the level of effort invested is also a significant positive predictor (i.e. higher effort leads to higher IM). The strength of individual predictors is practically independent of gender, however, girls expressed a significantly lower feeling of competence during practical work and felt greater pressure when experimenting. [...]

Finally, our research points out that a suitable level of demands and challenge is needed when designing practical work; low-challenging assignments can decrease intrinsic motivation mainly in students with a weaker, less positive relation to physics.

(Kácovský et al., 2023)

# conclusion

Our results show that after a 2-h programming workshop in a science museum, gender differences in self-efficacy for programming initially observed narrowed and even disappeared. Encouraging female participation in hands-on programming activities, even when those are of short duration, could be a particularly beneficial strategy to encourage girls' participation in computer science and to offer them exposure to mastery experiences. However, it remains unclear whether there would be lasting effect of a short workshop and if we would obtain similar findings with an older sample. Further experimental investigations are also needed to determine if single-sex setting was a determining factor in developing girls' self-efficacy beliefs for programming. Single-sex settings may allow girls to experience STEM without being overly exposed to strategies more frequently used by boys, such as dominance, competition, and suppression (Hoffmann, 2002). While the literature on women's experience in many STEM fields (physics, mathematics, biology, etc.) is quite abundant, few studies have explored how the gender composition of small group work affects younger girls' participation in programming. One interesting avenue would be to explore whether or not single-sex small group work is beneficial for girls' participation in programming activities and the development of their self-efficacy for programming.

(Allaire-Duquette et al., 2022)

This research allowed us to explore the development of an understanding of falling bodies that supports, adequately or not, the resolution of certain physics problems both in an atmosphere and in a vacuum. Using the prevalence framework, we have interpreted data extracted from the use of a cognitive task that recorded answers and latencies. Our conclusion led us to believe that while some misconceptions are quickly suppressed with age or formal instruction, others may be the result of schooling. High school and college science (physics) teachers seem to widely hold the belief that differently weighted objects will necessarily fall at exactly the same rate in atmospheric contexts, especially if their characteristics suggest that the friction (with the fluid through which they fall) will be equivalent for both objects. This belief is likely to result in teaching efforts that explicitly oppose the "naive" notion that mass is the essential key to solving falling body problems, to the no less naive attractor that mass never is; that falling objects will always accelerate at the same rate (9.8 m/s2), regardless of mass, or provided that friction is equivalent for all objects considered. Data from student participants seem to confirm this problem. Moreover, only university physics professors seem to avoid this pitfall.

We hope that our results will make physics-mechanics teachers aware of their possible adherence to, or misuse of, certain conceptual/misconceptual attractors. We also hope that researchers will see in our results yet another indication that didactic problems such as falling bodies should not be viewed as a succession of monolithic, isolated ideas to which adherence is exclusive and sequential, but rather as an evolving and organically woven web of conceptual resources that can nonetheless inform us about where to focus the subsequent innovation efforts.

(Potvin et al., 2023)

6. Conclusion

The aim of this research is to present a successful flipped classroom proposal in higher education to better understand its influence in terms of knowledge, skills and engagement. The reason why we focus on these three dimensions is due to their core roles in the international skills-oriented learning conceptual frameworks developed to enhance the employability of Generation Z students in the digital society of the twenty-first century. In doing so, first, we develop a flipped classroom measurement scale (4D_FLIPPED) to explore the degree of flipped classroom presence in our higher education learning experience. Second, we present a quantitative analysis by means of PLS-SEM to analyze the causal relationships of knowledge, skills, and engagement with students' satisfaction. To the best of our knowledge, this study is the first time that the two above contributions are accomplished in the literature. The empirical results point out that there are four fundamental dimensions that should be present in the flipped classroom to be successful in the 21st century with Generation Z. This study also confirms that the flipped classroom has positive effects on students' knowledge, skills, and engagement.

In terms of additional benefits not only to the literature but also to day-to-day practice, our research provides useful recommendations and insights for academia. Following the 4D_FLIPPED measurement scale tested in this investigation, course coordinators can consider how the flipped classroom can be incorporated into a learning design for their own courses in higher education. The purpose of this study is that our learning experience setup can be generalizable to other university contexts that might be interested in developing active and student-centered learning environments, as well as engagement and satisfaction generators, with the potential to acquire the knowledge and skills necessary to be successful in the workplace.

Finally, it is also worth mentioning that flipping the classroom implies increases in the workloads of both students and instructors. Complementary pedagogical approaches aimed at enhancing the engagement of students and instructors, such as gamification, crowdsourcing, digital transformation, and creativity, can help to render this workload more bearable. In this sense, future research could include the treatment of more complex frameworks by combining our flipped classroom proposal with these complementary pedagogical approaches in higher education. Additionally, the inclusion of variables such as time, gender, and language as different settings could provide further insights with reference to the model proposed in the present study.

(Murillo-Zamorano et al., 2019)

# références


Agogué, M., Poirel, N., Pineau, A., Houdé, O., & Cassotti, M. (2014). The impact of age and training on creativit : A design-theory approach to study fixation effects. *Thinking Skills and Creativity*, 11, 33-41. https://doi.org/10.1016/j.tsc.2013.10.002

Ahr, E., Houdé, O., & Borst, G. (2016). Inhibition of the mirror generalization process in reading in school-aged children. *Journal of experimental child psychology*, 145, 157-165.

Ahr, E., Houdé, O., & Borst, G. (2017). Predominance of lateral over vertical mirror errors in reading: A case for neuronal recycling and inhibition. *Brain and Cognition*, 116, 1-8.

Allaire-Duquette, G., Charland, P., & Riopel, M. (2014). At the very root of the development of interest : Using human body contexts to improve women's emotional engagement in introductory physics. *European Journal Of Physics Education*, 5(2), 31-48. https://doi.org/10.20308/ejpe.93516

Allaire-Duquette, G., Chastenay, P., Bouffard, T., Bélanger, S. A., Hernandez, O., Mahhou, M. A., Giroux, P., McMullin, S., & Desjarlais, E. (2022). Gender Differences in Self-efficacy for Programming Narrowed After a 2-h Science Museum Workshop. Canadian Journal of Science, *Mathematics and Technology Education*, 22(1), 87-100. https://doi.org/10.1007/s42330-022-00193-7

Ayotte-Beaudet, J.-P., & Potvin, P. (2020). Factors Related to Students' Perception of Learning During Outdoor Science Lessons in Schools' Immediate Surroundings. *Interdisciplinary Journal of Environmental and Science Education*, 16(2). https://doi.org/10.29333/ijese/7815

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow : Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 149(8), 1608-1613. https://doi.org/10.1037/xge0000729

Borst, G., Poirel, N., Pineau, A., Cassotti, M., & Houdé, O. (2013). Inhibitory control efficiency in a Piaget-like class-inclusion task in school-age children and adults : A developmental negative priming study. *Developmental Psychology*, 49(7), 1366-1374. https://doi.org/10.1037/a0029622

Brundiers, K., Barth, M., Cebrián, G., Cohen, M., Diaz, L., Doucette-Remington, S., Dripps, W., Habron, G., Harré, N., Jarchow, M., Losch, K., Michel, J., Mochizuki, Y., Rieckmann, M., Parnell, R., Walker, P., & Zint, M. (2021). Key competencies in sustainability in higher education—Toward an agreed-upon reference framework. *Sustainability Science*, 16(1), 13-29. https://doi.org/10.1007/s11625-020-00838-2

Burgoyne, A. P., Hambrick, D. Z., Moser, J. S., & Burt, S. A. (2018). Analysis of a mindset intervention. *Journal of Research in Personality*, 77, 21-30. https://doi.org/10.1016/j.jrp.2018.09.004

Chastenay, P., & Riopel, M. (2020). Development and validation of the moon phases concept inventory for middle school. *Physical Review Physics Education Research*, 16(2), 020107. https://doi.org/10.1103/PhysRevPhysEducRes.16.020107

Delalande, L., Moyon, M., Tissier, C., Dorriere, V., Guillois, B., Mevell, K., ... & Borst, G. (2020). Complex and subtle structural changes in prefrontal cortex induced by inhibitory control training from childhood to adolescence. *Developmental Science*, 23(4), e12898.

Eyme, K. M., Domin, M., Gerlach, F. H., Hosten, N., Schmidt, C. O., Gaser, C., Flöel, A., & Lotze, M. (2019). Physically active life style is associated with increased grey matter brain volume in a medial parieto-frontal network. *Behavioural Brain Research*, 359, 215-222. https://doi.org/10.1016/j.bbr.2018.10.042

# références


Hasni, A., & Potvin, P. (2015). Student's Interest in Science and Technology and its Relationships with Teaching Methods, Family Context and Self-Efficacy. *International Journal of Environmental & Science Education*, 10(3), 337-366.

Heilporn, G., Lakhal, S., & Bélisle, M. (2021). An examination of teachers' strategies to foster student engagement in blended learning in higher education. *International Journal of Educational Technology in Higher Education*, 18(1), 25.https://doi.org/10.1016/j.compedu.2017.09.011

Homer, B. D., Plass, J. L., Raffaele, C., Ober, T. M., & Ali, A. (2018). Improving high school students' executive functions through digital game play. *Computers & Education*, 117, 50-58.

Howell, A. A., Jordan, M., McKelvy, M., Wahi-Singh, B., & Shadmany, H. (2023). The science of science is fun : Assessing the impact of interactive science demonstrations through everyday experiences and near-peer role modeling. *International Journal of Science Education*, 45(5), 405-429. https://doi.org/10.1080/09500693.2022.2164473

Iwuanyanwu, P. (2022). What Students Gain by Learning Through Argumentation. International *Journal of Teaching and Learning in Higher Education*, 34(1), 97-107.

Jolles, D.D., van Buchem, M.A., Rombouts, S.A.R.B., & Crone, E.A. (2012). Practice effects in the developing brain: A pilot study. *Developmental Cognitive Neuroscience*, 2, S180-S191.

Kácovský, P., Snětinová, M., Chvál, M., Houfková, J., & Koupilová, Z. (2023). Predictors of students' intrinsic motivation during practical work in physics. *International Journal of Science Education*, 45(10), 806-826. https://doi.org/10.1080/09500693.2023.2175626

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training: Transfer of task-switching training. *Developmental Science*, 12(6), 978-990. https://doi.org/10.1111/j.1467-7687.2009.00846.x

Kramarski, B., Mevarech, Z. R., & Lieberman, A. (2001). Effects of Multilevel Versus Unilevel Metacognitive Training on Mathematical Reasoning. *The Journal of Educational Research*, 94(5), 292-300. https://doi.org/10.1080/00220670109598765

Loes, C. N., Culver, K. C., & Trolian, T. L. (2018). How Collaborative Learning Enhances Students' Openness to Diversity. *The Journal of Higher Education*, 89(6), 935-960. https://doi.org/10.1080/00221546.2018.1442638

Marec, C.-É., Tessier, C., Langlois, S., & Potvin, P. (2021). Change in Elementary School Teacher's Attitude Toward Teaching Science Following a Pairing Program. *Journal of Science Teacher Education*, 32(5), 500-517. https://doi.org/10.1080/1046560X.2020.1856540

Mevarech, Z., & Fridkin, S. (2006). The effects of IMPROVE on mathematical knowledge, mathematical reasoning and meta-cognition. *Metacognition and Learning*, 1(1), 85-97. https://doi.org/10.1007/s11409-006-6584-x

Murillo-Zamorano, L. R., López Sánchez, J. Á., & Godoy-Caballero, A. L. (2019). How the flipped classroom affects knowledge, skills, and engagement in higher education : Effects on students' satisfaction. *Computers & Education*, 141, 103608. https://doi.org/10.1016/j.compedu.2019.103608

# références


Özcan, Z. Ç., & Eren Gümüş, A. (2019). A modeling study to explain mathematical problem-solving performance through metacognition, self-efficacy, motivation, and anxiety. *Australian Journal of Education*, 63(1), 116-134. https://doi.org/10.1177/0004944119840073

Potvin, P., & Hasni, A. (2014). Analysis of the Decline in Interest Towards School Science and Technology from Grades 5 Through 11. *Journal of Science Education and Technology*, 23(6), 784-802. https://doi.org/10.1007/s10956-014-9512-x

Potvin, P., Masson, S., Lafortune, S. et Cyr, G. (2015a). Persistence of the intuitive conception that heavier objects sink more : a reaction time study with different levels of interference. *International Journal of Science and Mathematics Education*, 13(1), 21–34. http://dx.doi.org/10.1007/s10763-014-9520-6.

Potvin, P., Sauriol, É., & Riopel, M. (2015b). Experimental evidence of the superiority of the prevalence model of conceptual change over the classical models and repetition. *Journal of Research in Science Teaching*, 52(8), 1082-1108. https://doi.org/10.1002/tea.21235

Potvin, P., Hasni, A., Sy, O., & Riopel, M. (2018). Two Crucial Years of Science and Technology Schooling : A Longitudinal Study of the Major Influences on and Interactions Between Self-Concept, Interest, and the Intention to Pursue S&T. *Research in Science Education*, 50(5), 1739-1761. https://doi.org/10.1007/s11165-018-9751-6

Potvin, P., Malenfant-Robichaud, G., Cormier, C., & Masson, S. (2020a, September). Coexistence of misconceptions and scientific conceptions in chemistry professors: A mental chronometry and fMRI study. *In Frontiers in Education* (Vol. 5, p. 542458). Frontiers Media SA.

Potvin, P., Hasni, A., & Sy, O. (2020b). Attempting to Develop Secondary Student's Interest for Science and Technology Through an In-Service Teacher Training Initiative Based on the Principles of the Learning Community. Journal of Research in Science, *Mathematics and Technology Education*, 3(1), 15-34. https://doi.org/10.31756/jrsmte.312

Potvin, P., Chastenay, P., Thibault, F., Riopel, M., Ahr, E., & Brault Foisy, L.-M. (2023). An understanding of falling bodies across schooling and experience based on the conceptual prevalence framework. *Disciplinary and Interdisciplinary Science Education Research*, 5(1), 8. https://doi.org/10.1186/s43031-023-00075-4

Reinoso Tapia, R., Delgado-Iglesias, J., & Fernández, I. (2019). Learning difficulties, alternative conceptions and misconceptions of student teachers about respiratory physiology. *International Journal of Science Education*, 41(18), 2602-2625. https://doi.org/10.1080/09500693.2019.1690177

Remmen, K. B., & Frøyland, M. (2020). Students' use of observation in geology: towards 'scientific observation'in rock classification. *International journal of science education*, 42(1), 113-132.

Riopel, M., Nenciovici, L., Potvin, P., Chastenay, P., Charland, P., Sarrasin, J. B., & Masson, S. (2019). Impact of serious games on science learning achievement compared with more conventional instruction : An overview and a meta-analysis. *Studies in Science Education*, 55(2), 169-214. https://doi.org/10.1080/03057267.2019.1722420

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). *Interleaved practice improves mathematics learning. Journal of Educational Psychology*, 107(3), 900-908. https://doi.org/10.1037/edu0000001

# références

Sahin, D., & Yilmaz, R. M. (2020). The effect of Augmented Reality Technology on middle school students' achievements and attitudes towards science education. *Computers & Education*, 144, 103710. https://doi.org/10.1016/j.compedu.2019.103710

Schofield, L., Takriti, R., Rabbani, L., AlAmirah, I., Ioannidou, O., Alhosani, N., Elhoweris, H., & Erduran, S. (2023). Early years education teachers' perceptions of nature of science. *International Journal of Science Education*, 45(8), 613-635. https://doi.org/10.1080/09500693.2023.2168139

Tseng, A. S., Bonilla, S., & MacPherson, A. (2021). Fighting "bad science" in the information age: The effects of an intervention to stimulate evaluation and critique of false scientific claims. *Journal of Research in Science Teaching*, 58(8), 1152-1178. https://doi.org/10.1002/tea.21696