Guiding the minds of tomorrow : conversational agents to train curiosity and metacognition in young learners

Rania Abdelghani

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE
SOCIÉTÉS, POLITIQUE, SANTÉ PUBLIQUE (SP2)

SPÉCIALITÉ : SCIENCES COGNITIVES

Par Rania ABDELGHANI

Guider les esprits de demain: Agents Conversationnels pour Entraîner la Curiosité et la Métacognition chez les Jeunes Apprenants

sous la direction de
Hélène SAUZÉON et Pierre-Yves OUDEYER

Soutenue le : 03 / 09 / 2024

Membres du jury :

Mme ENGEL, Susan	Pr.	*Williams College*	Présidente
Mme ROMERO, Margarida	Pr.	*Université Côte d'Azur*	Rapporteuse
Mme RUGGERI, Azzurra	Pr.	*MPI Berlin, TU Munich & CEU Vienna*	Rapporteuse
Mme JIROUT, Jamie	PA	*Université de Virginie*	Examinatrice
M. VIE, Jill-Jênn	CR	*Inria Saclay & École Polytechnique*	Examinateur
Mme SAUZÉON, Hélène	Pr.	*Inria & Université de Bordeaux*	Superviseure
M. OUDEYER, Pierre-Yves	DR	*Inria Bordeaux*	Superviseur

Membres invités :

Mme De Vulpillières, Catherine	*EvidenceB*	Invitée

Thesis submitted for the degree of

Doctor of Philosophy of University of Bordeaux

Major in : Cognitive Science

Guiding the minds of tomorrow: Conversational Agents to Train Curiosity and Metacognition in Young Learners

By Rania ABDELGHANI

Supervision: Hélène SAUZÉON & Pierre-Yves OUDEYER

Defended on : 03 / 09 / 2024

Jury members:

Mrs. ENGEL, Susan	Pr.	*Williams College*	President
Mrs. ROMERO, Margarida	Pr.	*University of Côte d'Azur*	Reviewer
Mrs. RUGGERI, Azzurra	Pr.	*MPI Berlin, TU Munich & CEU Vienna*	Reviewer
Mrs. JIROUT, Jamie	PA	*University of Virginia*	Examiner
Mr. VIE, Jill-Jênn	CR	*Inria Saclay & École Polytechnique*	Examiner
Mme SAUZÉON, Hélène	Pr.	*Inria & University of Bordeaux*	Supervisor
M. OUDEYER, Pierre-Yves	DR	*Inria Bordeaux*	Supervisor

0 *Image generated by DALL·E with small edits.*
Prompt: "A lively classroom with teachers and artificial agents promoting curiosity, question-asking and exploration."

*T*here is no equality in education without social justice.

I dedicate this work to the memory of Hind, Reem, "The soul of my soul" and to that of all children forcibly taken away from us with war and violence.

And to those who stand no chance to access education and to nourish their curiosity. I hope, one day, research will also be for you ..

TITRE : GUIDER LES ESPRITS DE DEMAIN: AGENTS CONVERSATIONNELS POUR ENTRAÎNER LA CURIOSITÉ ET LA MÉTACOGNITION CHEZ LES JEUNES APPRENANTS

RESUMÉ

La curiosité épistémique (CE), i.e. le désir d'explorer une information pour le plaisir qu'elle procure, est un phénomène étudié dans divers domaines. Plusieurs chercheurs ont souligné son rôle fondamental dans le développement cognitif et la promotion d'un apprentissage continu. De plus, la CE est considérée comme clé pour cultiver un esprit capable de s'adapter aux incertitudes du monde.

Ces recherches ont suscité un grand intérêt pour la CE en éducation, la considérant essentielle pour permettre aux individus d'être actifs et maîtres de leur apprentissage. Ce sont des propriétés cruciales pour relever certains des défis éducatifs: offrir aux élèves un soutien adapté à leurs compétences et motivations, et les aider à être des apprenants autonomes et indépendants dans des environnements dynamiques et incertains.

Malgré son importance, l'implémentation de la CE dans les salles de classe reste limitée. Notamment, l'une des principales expressions de la CE—le questionnement— est presque absente dans la plupart des établissements: les élèves sont souvent amenés à répondre aux questions des enseignants plutôt qu'à poser les leurs. Et lorsqu'ils posent des questions, elles sont généralement de bas niveau et, contrairement aux questions curieuses, ne cherchent pas de nouvelles informations majorantes aux connaissances antérieures.

Cette thèse propose donc de développer des technologies éducatives qui visent à favoriser l'apprentissage dirigé par la CE, en entraînant les comportements de questionnement curieux et les compétences qui lui sont liées. Pour cela, nous proposons des interventions pour entraîner trois dimensions:**1) Les compétences linguistiques de questionnement:** On implémente un agent conversationnel pour aider les élèves à générer des questions curieuses lors de tâches de lecture-compréhension. L'agent fournit des indices spécifiques pour faciliter l'utilisation des mots interrogatifs composés et des constructions interrogatives. Différentes structures d'indices (phrase vs. série de mots-clés) et leurs modes de génération (manuellement vs. par GPT-3) sont étudiées. **2) Les compétences métacognitives (MC) liées à la CE:** On crée des vidéos animées pour donner des connaissances

déclaratives sur les compétences MC liées à la CE: l'autoréflexion, faire des hypothèses, formuler des questions et évaluer les nouvelles informations. On propose également des sessions pour pratiquer ces compétences lors de tâches de lecture-compréhension, en utilisant des indices donnés par des agents conversationnels conçus pour entraîner la MC procédurale. **3) Les perceptions sociales:** On crée des vidéos animées pour expliquer la CE et sa mise en pratique pour corriger les idées négatives qu'ont les apprenants sur ce concept.

Plus de 150 élèves français âgés de 9 à 11 ans ont été recrutés pour tester l'entraînement de ces dimensions. Combinées, ces dernières ont amélioré la sensibilité MC des élèves et leur perception de la curiosité. Ces deux facteurs ont, à leur tout, facilité les comportements de questionnement divergent. Cela a également conduit à un progrès d'apprentissage plus fort et à des expériences d'apprentissage positives et soutenables. Mais malgré ces résultats, nos méthodes présentent certaines limites, en particulier leur courte durée. Cette thèse encourage donc le travail sur des solutions plus durables afin d'examiner les effets à long terme sur la CE.

Enfin, cette thèse souligne la nécessité de continuer à explorer les recherches sur le questionnement et la MC à l'âge de l'intelligence artificielle générative (IAG). Bien que la IAG facilite l'accès à l'information, elle nécessite encore de bonnes capacités de questionnement et de MC, pour prévenir les mésuages et/ou faciliter leur détection. Nous proposons un Framework liant l'utilisation efficace de la IAG en éducation, les compétences de questionnement et de MC, et la littératie en IAG. Nous présentons également une étude comportementale pour tester ces relations.

Mots Clés: Curiosité épistmique / métacognition / technologies éducatives / intelligence artificielle en éducation / Questionnement / IA Génératives / agents conversationnels

TITLE: GUIDING THE MINDS OF TOMORROW: CONVERSATIONAL AGENTS TO TRAIN CURIOSITY AND METACOGNITION IN YOUNG LEARNERS

ABSTRACT

Epistemic curiosity—the desire to actively seek information for its inherent pleasure—is a complex phenomenon extensively studied across various domains. Several researchers in psychology, neuroscience, and computer science have repeatedly highlighted its foundational role in cognitive development and in fostering lifelong learning. Further, epistemic curiosity is considered key for cultivating a flexible mindset capable of adapting to the world's uncertainties.

These insights have spurred significant interest in the educational field, recognizing curiosity as essential for helping individuals be active and in control of their learning. These properties are crucial for addressing some of today's major educational challenges, namely offering students individualized support to suit their competencies and motivations, and helping them become able to learn autonomously and independently in their dynamic and uncertain environments.

Despite this well-documented importance of curiosity in education, its practical implementation and promotion in the classroom remains limited. Notably, one of the primary expressions of curiosity—question-asking (QA)—is nearly absent in most of today's educational settings. Several reports show that students often spend a lot of time answering teachers' questions rather than asking their own. And when they do ask questions, they are typically low-level and memory-based, as opposed to curious questions that seek novel information.

In this context, this thesis aims to develop educational technologies that can foster children's curiosity-driven learning by practicing curious QA behaviors, and their related metacognitive (MC) skills. Ultimately, we implemented interventions to train three dimensions: **1) Linguistic QA Skills:** We implement a conversational agent to train the ability to formulate curious questions using compound questioning words and correct interrogative constructions. It helps children generate curious questions during reading-comprehension tasks, by providing specific cues. The effectiveness of different cue structures (a sentence vs. series of keywords) and implementations (hand-generated vs. GPT-3-generated content) is studied. **2) Curiosity-related metacognitive Skills:** We create animated videos to give declarative knowledge about curiosity and its related MC skills: the ability to self-reflect, make educated guesses, formulate efficient questions, and evaluate newly-acquired information. We also propose sessions to practice these skills during reading-comprehension tasks using specific cues given by conversational agents we designed to train procedural MC. **3) Social Perceptions and beliefs:** We create animated videos to address the negative constructs learners tend to have about curiosity. They explain the importance of curiosity and how to control it during learning.

Over 150 French students aged 9 to 11 were recruited to test these trainings of the three dimensions. Combined, these latter enhanced students' MC sensitivity and perception of curiosity. At their turn, these factors facilitated students' divergent QA behaviors which, at their turn, led to stronger learning progress and positive, affordable learning experiences. But despite the positive results, our methods had limitations, particularly their short duration. We suggest testing longer-lasting interventions to examine their long-term effects on curiosity.

Finally, this thesis highlights the need to continue exploring QA and MC research in the age of Generative Artificial Intelligence (GAI). Indeed, while GAI facilitates access to information, it still requires good QA abilities and MC monitoring to prevent misinformation and facilitate its detection. We thus propose a framework to link efficient GAI use in education to QA and MC skills, and GAI literacy. We also present a behavioral study we intend to conduct to test this framework.

Key words: Curiosity / Metacognition / Educational technologies / AI in education / Question-asking / generative AI / conversational agents

---

Équipe-projet FLOWERS, Centre INRIA de l'Université de Bordeaux. 200 Av. de la Vieille Tour, 33405 Talence

PUBLICATIONS

Some of the work presented in this thesis has appeared in the following publications:

- Rania Abdelghani et al. "Conversational agents for fostering curiosity-driven learning in children." In: *International Journal of Human-Computer Studies* 167 (2022), p. 102887. URL: https://www.sciencedirect.com/science/article/abs/pii/S1071581922001112
- Rania Abdelghani et al. "Gpt-3-driven pedagogical agents to train children's curious question-asking skills." In: *International Journal of Artificial Intelligence in Education* (2023), pp. 1–36. URL: \url{https://link.springer.com/article/10.1007/s40593-023-00340-7}
- Rania Abdelghani et al. "Generative AI in the Classroom: Can Students Remain Active Learners?" In: *arXiv preprint* (2023). URL: https://arxiv.org/abs/2310.03192
- Rania Abdelghani et al. "Interactive environments for training children's curiosity through the practice of metacognitive skills: a pilot study." In: *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 2023, pp. 495–501. URL: https://dl.acm.org/doi/abs/10.1145/3585088.3593880
- Xingdi Yuan et al. "Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 12952–12965. URL: https://aclanthology.org/2023.findings-acl.820
- Ziang Xiao et al. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding." In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023, pp. 75–78. URL: https://dl.acm.org/doi/abs/10.1145/3581754.3584136

ACKNOWLEDGEMENTS

My intellectual journey has been filled with ups and downs, moments of confidence and doubt, of relief and frustration. Going through this roller coaster of emotions while away from my family and friends will probably always be one of the toughest challenges I have to take on. So today, looking back at these three and a half hectic years, I can only be over the moon with gratitude for everyone who supported me through all of this. Thank you to my family and friends who were continuously sending love and support from all over the world, to my academic peers and the lovely people I met in Bordeaux.

First, my deepest gratitude goes to all of the wonderful women of colour in academia. I found a lot of courage and inspiration looking at your achievements. Especially as someone struggling to find a place in academia while keeping responsibilities toward my community and my cultural heritage as my compass. I am grateful for you all !

Thank you to my supervisors Hélène Sauzéon and Pierre-Yves Oudeyer for their time and guidance throughout these years. Thank you for letting me make mistakes and for valuing my work.

Thank you to all the teachers who took of their time to help us run our studies despite their busy schedules. A huge gratitude to you all, none of all this work could have happened without your precious help. Thank you Mr. Stéphane Théas, Mme. Charlotte Fourcade, Mme. Cynthia Laroche, Mme. Rachel Pomès, Mme. Isabelle Coutand, Mme. Mélanie Laurent, and Mme. Justine Ardennes. Thank you also to Mr. Didier Roy who helped us connect with schools and make all this happen. And of course, thank you to all the children that have participated in the studies and helped me understand a lot of things. Beyond collecting data, I really enjoyed visiting the schools and meeting all the wonderful children. I do hope that our presence was somewhat enjoyable and beneficial.

Thank you to my awesome lab-mates, you really are exceptional and made long hours at the office bearable. Thank you Gautier for the burn-out pauses and the ping-pong battles (you really need to change your music taste though). Thank you Marion, Isabeau, Matisse, Thomas, Maxime for the coffee pauses and cool heart-filling discussions. Thank you Grgur, Clément, Mayalen, Laetitia for welcoming me to the team and to Bordeaux, it was nice to meet all of you. Thank you to Cécile also for always bringing the smile and being available when we need you. Thank you also to Chloé, Jérémy, Julien and Corentin, the newest 'Flowers'. I hope your time at the lab will be as awesome as mine. Thank you also to Karim for bringing the positive energy at my late evening stays at the lab. And finally, thank you to my colleagues at EvidenceB, it was a pleasure to work with such a nice group of committed and passionate people.

Thank you also to the "Kidd Lab" for welcoming me for a few months. It was a pleasure to spend time in such a vibrant and dynamic research environment. I am grateful for the opportunity. Special thanks to Pr. Celeste Kidd, Holly Palmeri and my supervisors for helping me with this opportunity.

Thank you to all my friends who were always there for me even when I am oceans away. I don't know where to start so here's a random rambling. Poupou, Mejoul, Sarhan, Mariem, Fifi, Ous, Jean, Bech, Anne-Claire. Adulthood has been hard without you! I will always cherish the times we spent laughing and being crazy, angry and high on life.

Thank you to my mom, my cornerstone who taught me love, hard work and resilience. And to my sweet dad who was always there for me without complaining. Thank you for always reminding me that it is okay to be sensitive and to take time for myself. I love you both so much and I will always be so sad for not having as many chances as I want to say this to you in person.

To my brother. I honestly don't think I could have been here today without you. Whenever life gets complicated, I always find a way out because I know that you will be there to help. I hope you get all the happiness you deserve. I also hope all girls can have a big brother like you.

And finally, to my partner. I know I've been insupportable during these past months (in normal times too, but I think these times have been worse). Huge thank you for you patience and support throughout these challenging times. Thank you for always encouraging me and for being my family in this city.

Cheers to everyone who has been there for me, supported me in any way and has been kind to me in these 28 years !

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

EC Epistemic Curiosity

QA Question-Asking

MC Metacognition

LP Learning Progress

IM Intrinsic Motivation

KG Knowledge Gap

ZPD Zone of Proximal Development

AI Artificial Intelligence

GAI Generative Artificial Intelligence

NLP Natural Language Processing

EdTech Educational Technology

CA Conversational Agent

ITS Intelligent Tutoring System

LLMs Large Language Models

CIAC Children's Images and Attitudes of Curiosity Questionnaire

MSLQ Motivated Learning Strategies Questionnaire

NasaTLX Nasa task-load Questionnaire

Part I

INTRODUCTION

1 THESIS OVERVIEW

1.1 GENERAL INTRODUCTION

Human learning heavily relies on the ability to make sense of the world's dynamic constraints and adapt behavior to achieve specific goals and standards. From a very young age, humans can discover and learn basic relationships in their environment, such as simple physical laws, through mere observation. As they develop, they become capable of understanding even more complex phenomena, such as statistical patterns and the psychology and emotions of others. For this, humans do not rely solely on observation; instead, they actively interact with the environment, make inferences and conduct carefully designed explorations and experiments.

But what drives these behaviors? How do individuals decide when they need to seek knowledge? What is it that motivates them to seek knowledge even at the expense of their safety and comfort sometimes? One initial explanation is that humans pursue knowledge they believe will fulfill their specific needs and bring them happiness or satisfaction [23]. These needs may be driven by the anticipated material benefits of acquiring the knowledge in question. For example, experimenting with different types of sweets—despite the risk of encountering ones we dislike—can be gratifying because it leads to discovering a preferred candy, thereby enhancing its enjoyment. However, the pursuit of knowledge can also be "non-instrumental," where the motivation stems solely from the satisfaction of learning. In this case, the only reward is the information itself. This phenomenon, widely known as intrinsically-motivated information-seeking, characterizes individuals when they are being *curious*.

Despite being fundamental to human behavior, much about curiosity is still unknown: its origins, triggering mechanisms, etc. Such questions have long captivated scientists across various disciplines, sparking extensive research in human behavior, psychology, neuroscience, education, artificial intelligence, etc. But even after decades of curiosity research, consensus on its precise nature and mechanisms remains elusive. Theories range from classical perspectives that view curiosity as a fundamental drive motivating individuals to seek novel stimuli (e.g. [90, 221]), to more nuanced explanations presenting it as a mechanism to help maximize learning potential by reducing uncertainty about the world (e.g. [55, 147, 176]).

While these propositions vary, most of them share a common underlying idea: individuals interact with their environment and receive information from it. The difference between this new information and their existing knowledge can create an informational need—a learning goal aimed at bridging the gap between the current knowledge state and a desired one. This need can motivate subsequent information-seeking behaviors, or what we call *curiosity-driven learning*. In this process, individuals evaluate their informational progress during exploration, compare it to their initial goals, and maintain their behavior until they feel satiated.

This idea represents a significant advancement in modeling the mechanisms of curiosity. However, it presupposes various skills necessary for engaging in curiosity-driven exploration, such as integrating novel information into existing knowledge, using this integration to deduce uncertainties about the world, monitoring progress to adapt the search process, etc [168]. These skills are metacognitive in nature: they involve an understanding of one's own knowledge and the ability to monitor its progress. Numerous studies support indeed this idea of curiosity as a metacognitive feeling/ experience, suggesting that it arises from a complex interplay between cognitive factors (such as already-acquired knowledge about a topic) and metacognitive functions [77]. But despite these compelling suggestions, other research disagrees, viewing curiosity as a simple motivational experience that may or may not involve metacognition [39]. See Figure 1.1 for a proposition of the role of metacognition in curiosity-driven information-search.

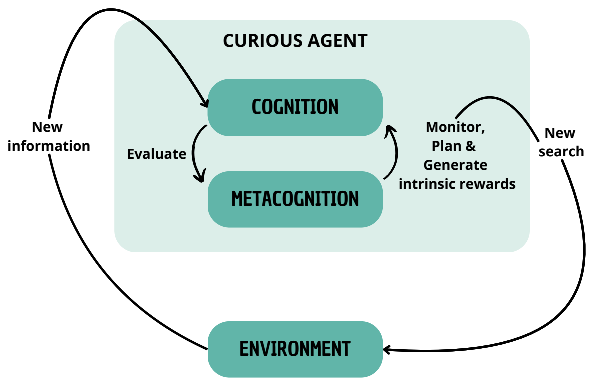


Figure 1.1: A proposition for the role of metacognition in curiosity-driven information-search

While we still lack consensus on a clear and unique definition of curiosity, research has converged on its foundational role in enhancing learning by improving memory, control, and agency [117]. Consequently, educational research has shown keen interest in curiosity, recognizing it as a catalyst for engaging students in tasks aligned with their interests, competency levels, and learning goals. This is particularly important in modern education, where the world's challenges and uncertainties abound, and individuals exhibit increasingly diverse capabilities and motivations.

However, despite its importance in learning, reports continue to highlight a lack of curiosity-driven behaviors among children in classrooms, as well as a deficiency in curiosity-driven teaching approaches. One key behavior of interest is question-asking, often considered as a primary behavioral manifestation of curiosity. Indeed, although questions can have different functions, their primary function remains epistemic: acquire specific new information and expand one's knowledge. In classrooms, students are often found lacking this behavior. Instead, they are typically asked to respond to their teachers' questions to demonstrate their knowledge. And even when they do ask questions, they are usually low-level and memory-based (e.g. seeking confirmation), as opposed to curious questions that seek novel information.

Research investigating the absence of curiosity-driven behaviors in classrooms offers several explanations. They include social factors (e.g., fear of negative judgment when asking questions) [103], linguistic barriers (i.e., difficulties in formulating well-structured questions that accurately convey informational needs) [69], lack of motivation and sense of agency [223], and importantly, deficits in students' metacognitive skills. This means that students may refrain from asking questions because they fail to recognize their need for specific information. This tendency to overestimate one's own knowledge, often referred to as "the knowledge illusion," is frequently observed in younger students and is attributed, among other factors, to their metacognitive immaturity [129].

In this context, the overall goal of this thesis is to use our knowledge about curiosity mechanisms to propose methods for enhancing curiosity-driven question-asking behaviors and their related skills. We focus on training three dimensions: 1) linguistic QA skills to facilitate the use of high-level questioning words and interrogative constructions. 2) Curiosity-related metacognitive Skills to facilitate the ability to self-reflect, make educated guesses, formulate efficient questions, and evaluate newly-acquired information. And 3) Social Perceptions and beliefs to address the negative constructs learners may have about curiosity. We leverage emerging technologies such as conversational agents and natural language processing methods (NLP), including Generative Artificial Intelligence (GAI), to design and implement pedagogical trainings for these dimensions. Indeed, we hypothesize that, when used responsibly, GAI-based methods can offer several advantages in facilitating the design and implementation of curiosity-driven pedagogical activities in the classroom. Mainly, they can reduce the load of time-consuming tasks such as generating the pedagogical content for the proposed activities.

Ultimately, this thesis had three contributions. First, we design an expert-scripted conversational agents to help train divergent QA skills for 9-to-11 year old students— age when they start learning about QA constructions at school and forming negative perceptions towards curiosity [183]. We evaluate their effectiveness on exploratory behavior and learning outcomes. Second, we explore the feasibility of using GAI to generate the content for a divergent question-asking training and compare it to hand-scripted ones. Lastly, we create animated videos and design conversational agents to give a declarative and experience-based training of curiosity and its related metacognitive skills. We study the impact on children's spontaneous curiosity-driven QA behaviors.

To develop the material for these trainings and our testing methods, we rely on collaborations with primary-school teachers to help us validate their pedagogical relevance and accessibility. We organized several co-working sessions to collect their feedback about our propositions prior to starting the interventions. Once pedagogically validated, we then started testing the trainings with their students in their real classroom settings. See Figure 1.2 for an overview of our methods and procedures.

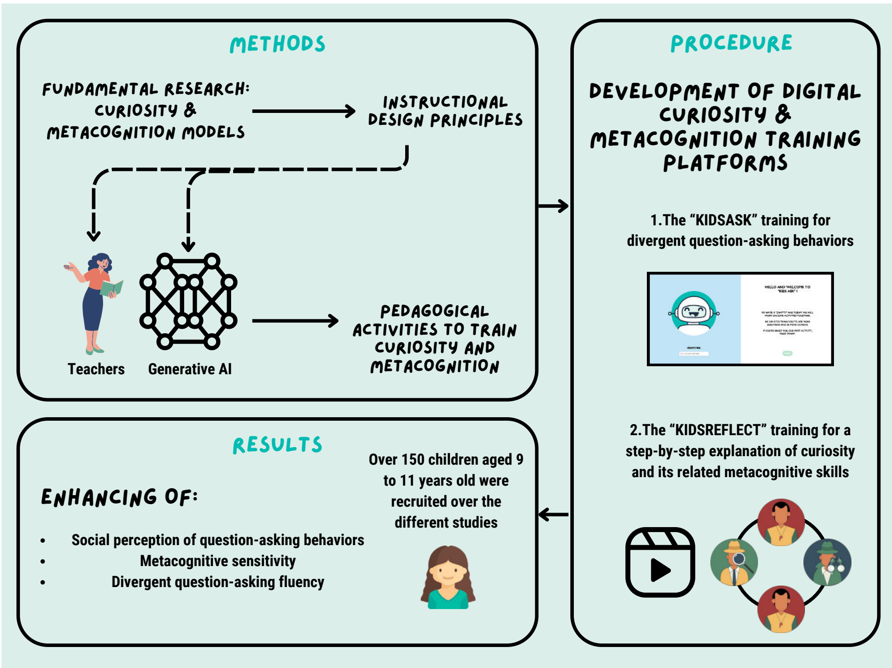


Figure 1.2: Ph.D. Methods and Contributions overview

On a final note, this thesis explores the open questions surrounding the role of curiosity and metacognition in the age of GAI. It proposes a framework and a behavioral study to investigate the link between these skills and an informed use of GAI in education.

1.2 METHODS AND CONTRIBUTIONS

Our approach relies on methods in educational psychology, human-centered design, human-AI interactions and behavioral experiments, drawing inspiration from literature on epistemic curiosity, metacognition and their promotion using learning technologies.

1.2.1 *Literature review*

In Chapter 2, we begin by reviewing the diverse literature on curiosity mechanisms, their connection to metacognition, and the pedagogical strategies for fostering curiosity-driven learning and related metacognitive strategies in the classroom. Additionally, we explore literature on educational technologies, including digital tools in general and NLP-based tools more specifically, to understand their potential in promoting curiosity and metacognitive skills.

In this chapter, we initially delve into the literature on intrinsic motivation and curiosity, aiming to provide an overview of the various existing explanations and models of curiosity. We identify internal and external factors that can influence curiosity, and then review its functions and roles in cognitive development and learning. Additionally, we use this review to gain insights into a key curiosity-driven behavior central to this thesis: divergent question-asking, i.e. asking questions that seek novel information related to the task at hand. This includes examining the taxonomy of questions, identifying those most closely associated with curiosity, and understanding their developmental trajectory and role in enhancing learning.

Second, we explore the literature related to metacognition, tracing its development from basic forms observed during infancy to more complex manifestations such as self-reflection in later stages of life. We then investigate the various links between metacognition and curiosity, focusing on how metacognitive skills can facilitate and support curiosity-driven learning. Specifically, we explore the idea that metacognitive abilities can promote curiosity-driven information-seeking by enabling individuals to recognize gaps in their knowledge and to plan and monitor their learning progress as they work to fill those gaps.

Finally, we discuss how research findings in curiosity and metacognition can be applied in the classroom to enhance students' learning experiences and outcomes. We explore various teaching strategies and approaches that can support and encourage these skills, such as promoting inquiry-based and puzzling activities over traditional instruction-based methods. Additionally, we address the challenges associated with implementing such strategies, including the time required to generate pedagogical content and the need to adapt to students' diverse learning preferences, interests, and motivations. Furthermore, we explore the potential benefits of using digital tools to address these challenges, particularly by leveraging new advances in artificial intelligence and natural language processing. Throughout this discussion, we identify studies that have used such techniques to implement curiosity and metacognition training, gaining insights into their potential pedagogical efficacy and enhancement of curiosity-driven QA behaviors.

1.2.2 *Empirical contributions*

After setting the landscape for our work, we present the empirical contributions of the thesis in Part iii. We report on three behavioral experiments involving pedagogical interventions aimed at enhancing curiosity-driven QA behaviors and related metacognitive strategies in students aged 9 to 11 years in Bordeaux, France. These trainings used digital platforms with activities designed based on our understanding of curiosity and metacognition theories and links. We relied on using conversational agents to administer these trainings and collaborated with teachers to validate their pedagogical relevance prior to launching our studies.

1.2.2.1 *Designing, implementing and evaluating "KidsAsk": a platform for training divergent QA skills*

During the first study (presented in details in Chapter 3), our aim was to develop a platform, we call "KidsAsk", offering interactions with a CA that can help children generate curiosity-driven questions during reading-comprehension tasks. The agent's strategy is to highlight potential learning opportunities that can spark curiosity about for the task (i.e. novel information about the text, in the form of a short sentence at hand, that can help deepen their understanding of it). Once children choose a learning opportunity proposition, the agent prompts them to formulate the question that can help solve it. It assists them in this process by providing an additional linguistic cue: a high-level questioning word that, combined with the earlier semantic cue, guides them to generate a divergent question. See Figure 1.3 for an overview of the platform.

The rationale behind our agent's behavior is two-fold: 1) it gives propositions of learning goals (i.e. missing information about the text) in line with curiosity models suggesting that individuals are mostly motivated to ask questions when they realize that they are missing out on some specific information. For developmental reasons, children tend to overestimate their knowledge and fail to identify these gaps. The CA is thus designed to give them this push in order to facilitate QA. The agent also gives linguistic cues as we hypothesize that syntactic challenges can prevent children from asking high-level questions.

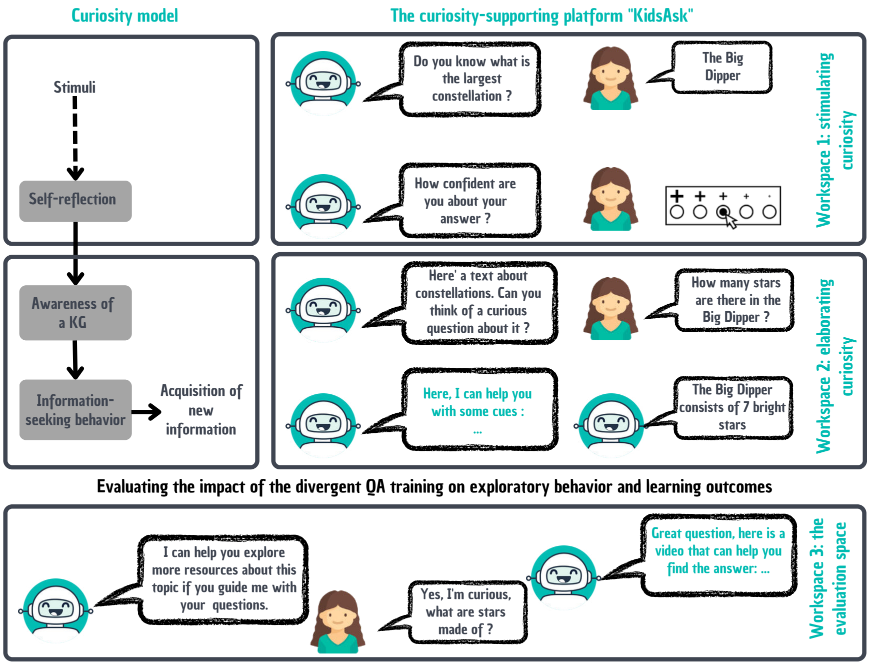


Figure 1.3: An overview of the "KidsAsk" training platform design principles and rationale

To evaluate the efficiency of this "incentive" agent, we recruited 51 primary school students aged between 9 and 11. They interacted with either the "incentive" agent, which provides both learning goal propositions and linguistic help, or a "neutral" agent that only offers linguistic assistance without highlighting knowledge gaps. Our data analysis focused on the percentage of divergent questions generated by children, both during the training and during another exploration-based activity children had to perform afterwards (see Figure 1.3). To do this, We manually annotated the questions based on their divergence level and syntactic quality using standardized grids. Additionally, we measured spontaneous exploratory behaviors and domain-knowledge learning progress and investigated their link with the participants' divergent QA behaviors.

Our results show that participants were more capable of generating divergent questions with the incentive agent compared to the neutral one. These findings persisted both during the training sessions (with the semantic and linguistic help of the CA) and during a subsequent activity (with no external help or incentives to formulate questions). This suggests a transfer effect of the divergent QA training to new educational contexts. Consistent with theories highlighting the role of curiosity in learning, we also find positive correlations between QA performance and both spontaneous exploratory behaviors and learning progress.

1.2.2.2 *Using LLMs to generate the pedagogical content for the "KidsAsk" training*

For the second study (presented in details in Chapter 4), we try to address two main limitations we saw in the first one: 1) generating all the pedagogical content for the "incentive" agent (i.e. the different propositions of knowledge gaps for every new text and the linguistic cues) was very time-consuming. This makes scaling up our approach and using it for other pedagogical activities quite challenging. And 2) participants were provided with predefined propositions of knowledge gaps, leading to specific questions that the *pedagogical teams* deemed important for children to ask. However, this heavily expert-directed approach might provide cues that do not align with some children's interests and/or competency levels, potentially hindering their curiosity. As a result, we suggest that "KidsAsk" may have primarily targeted the linguistic skills that facilitate curiosity-driven QA rather than fostering curiosity-driven QA behaviors themselves.

To work around these two main limitations, we propose two new implementations of our CA: 1) To tackle the up-scaling problem„ we propose an agent that uses recent advances in Large Language Models (LLMs) to generate the content of the training. This agent is designed to mimic the exact behavior of our "incentive" agent described the first study, we call it the *automated incentive agent*. 2) To address the restrictive nature of the propositions given by the incentive agents, we propose another LLM-based agent to generate the content of the QA training. However, this agent proposes a new structure for its propositions of knowledge gaps: it gives 'open' propositions. These open propositions lead children to imagine various questions, allowing them to generate questions that align with their interests. We call this agent the *automated open agent*.

To ensure appropriate interactions, we implemented a human-in-the-loop approach. This involved first checking the soundness of the content generated by GPT-3 and only connecting it to the agent for interaction with the children once it was validated. See Figure 1.4 for the behaviors of the different agents.

By comparing the expert-based incentive agent and the automated one, we aim to investigate the efficiency of using a LLM (GPT-3) to generate the pedagogical content of a divergent QA training. By comparing these two with the open automated agent, we explore whether having a more 'open' structure of the cues can help children generate more curiosity-driven questions.

To perform these comparisons, we recruited 75 students aged between 9 and 11 who interacted with either one of the three CAs. We performed our data analysis in two steps: first, we analyzed the quality of the content generated by GPT-3 both semantically (semantic relatedness to the texts, divergence level) and syntactically. Second,

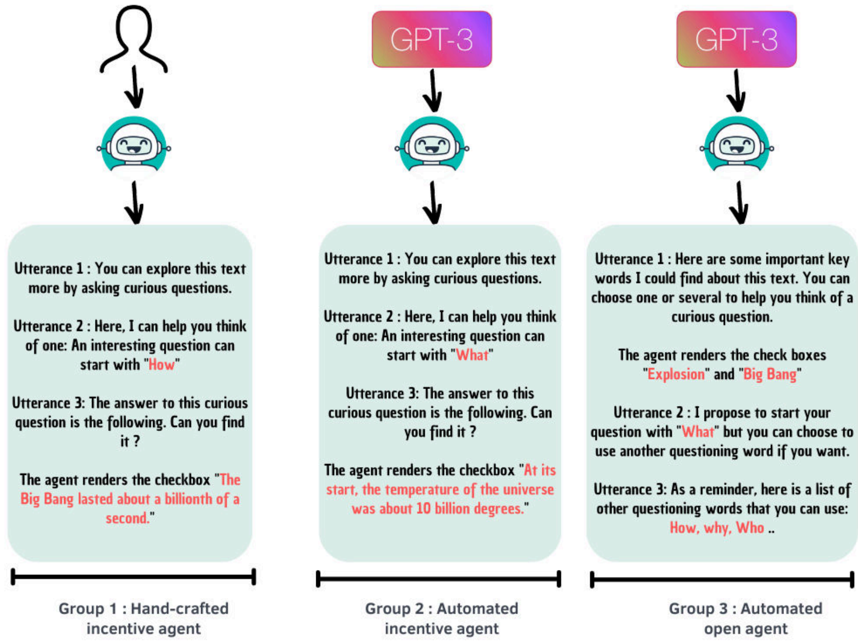


Figure 1.4: Implementations of "KidsAsk" using GPT-3

we compare the impact of the training in its three forms on children's divergent QA behaviors. Similar to what we did in the first study, we use manual annotation to determine the divergent level of a question.

Our results indicated that the quality of the content generated by GPT-3 was very similar to that generated by experts. Additionally, we saw similar behaviors for children who had the expert-based and the automated incentive agents. However, we saw a significantly better performance for those who interacted with the open one. Furthermore, this latter group was the only one where participants' QA behaviors correlated with their curiosity trait as reported by their parents (the more they are curious by trait, the more they asked divergent questions using the agent's prompts). Taken together, these results suggest the efficiency of GPT-3 to facilitate generating curiosity-prompting pedagogical content and that offering 'open', student-directed cues can help support curiosity better than the incentive and teacher-guided ones.

1.2.2.3 *Designing, implementing and evaluating "KidsReflect": a training for curiosity perception and curiosity-related metacognitive skills*

Finally, in our third study (presented in details in Chapter 5) we try to address the role of curiosity perceptions and metacognition in promoting children's *spontaneous* divergent QA behaviors. Indeed, "KidsAsk" primarily focuses on helping children find a relevant QA syntax formulation to compensate for a knowledge gap, but only once this latter is identified using the help of an external agent—the CA. However, a crucial preliminary step is identifying an intriguing uncertainty or a knowledge gap and to be motivated to pursue it. This means that asking curiosity-driven questions requires strong metacognitive skills to be able to identify this missing knowledge: both declarative metacognitive knowledge (i.e. knowledge about one's own knowledge, strategies and task at hand) and procedural metacognitive knowledge (i.e. the ability to apply specific learning strategies and to control, monitor them) are thus essential. Finally, positive perceptions of curiosity and QA are also crucial in order to be motivated to work towards compensating a knowledge gap once identified. With this bigger picture in mind, it can be arguable that "KidsAsk" only addresses a part of the procedural MC training.

Consequently, our goal for this third study is to design an intervention to help children be more efficient and comfortable with recognizing learning opportunities and with adopting strategies such as QA to resolve them. To do this, the intervention explains curiosity and its functions to correct potential negative perceptions, then proposes a training for both declarative and procedural metacognitive knowledge. We call this intervention "KidsReflect".

"KidsReflect" has two main components: 1) part I starts by introducing children to curiosity and its importance to have an enjoyable and efficient learning experience. We do this using animated videos we created within the team. The videos then introduce declarative knowledge about metacognition with the aim to give children conceptual knowledge about how to understand their own knowledge and functioning. More particularly, the videos presented declarative knowledge about how to: select a curious learning goal given one's previous knowledge and the task at hand, to estimate the expected knowledge gain before pursuing a specific information (by making educated guesses), to adapt the relevant information-search behavior to achieve the selected goal and finally, to evaluate the result of the search process and decide on subsequent behavior. To identify these MC skills, we took inspiration from the curiosity-driven autonomous learning framework proposed in [168] and attempted to operationalize it by linking its different components to specific metacognitive skills we hypothesize are essential for their achievement.

And 2) part II of "KidsReflect" proposes to give procedural/ experience-based knowledge about metacognition by putting these four skills into practice during a reading-comprehension task. The main goal of this task being to practice the control and monitoring of one's learning strategies in order to independently gain novel information about the text. To do this, we implement a digital platform with hand-scripted CAs that act as learning companions during the task and help children practice the skills described above by proposing specific hints and explanations. See Figure 1.5 for an illustration of "KidsReflect".

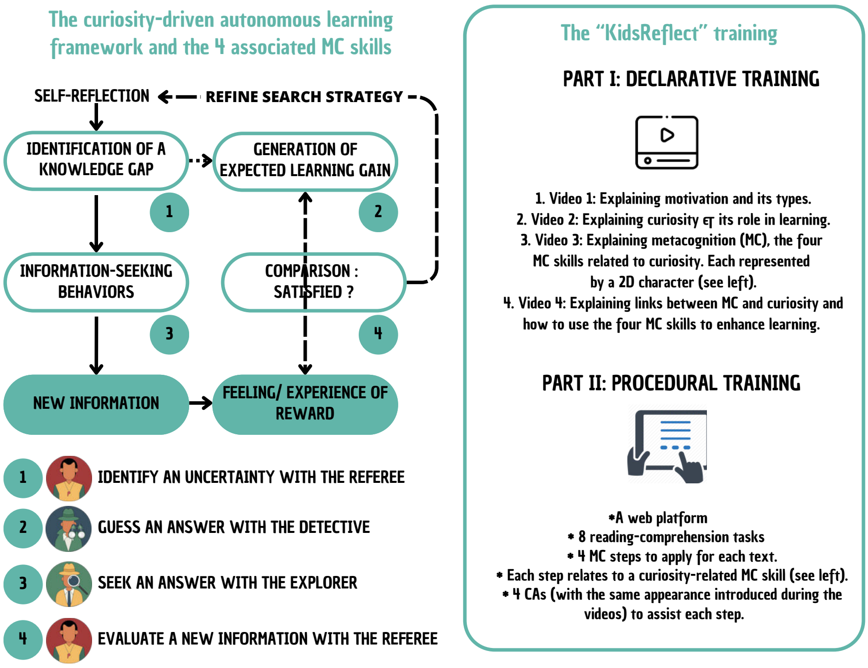


Figure 1.5: An overview of the "KidsReflect" training design rationale and content

To test this training, we recruited 86 students aged between 9 and 11, that were assigned to one of the three groups: "KidsAsk" only, part I only of "KidsReflect" + "KidsAsk" or the full "KidsReflect" + "KidsAsk". By comparing these groups, we aim to understand two key points: 1) the contribution of a training that targets social perceptions of curiosity and its related metacognitive skills on children's spontaneous curious QA behaviors. And 2) whether video-based explanations alone can already help enhance such behaviors. This could be an interesting finding as videos can be easily appropriated by teachers and introduced on a large scale in different schools.

Our results show indeed a significant impact of "KidsReflect" on children's metacognitive sensitivity (i.e. their accuracy in judging their knowledge level and performance) and their perception of QA behaviors in the classroom. At their turn, these two indicators, combined with children's performance during "KidsAsk" and their curiosity trait, showed to be in direct correlation with the spontaneous offline divergent QA behaviors. Taken together, these findings represent an exciting step in defining the proper strategies to put in place in order to foster children's curious QA behaviors in the classroom. These strategies include: practicing children's evaluative and monitoring metacognitive skills, enhancing their QA linguistic skills, and establishing positive views of curiosity. Such strategies could be adopted by pedagogical teams and educational technologies designers to better support curiosity-driven learning.

1.3 GENERAL DISCUSSION, PERSPECTIVES AND OPEN QUESTIONS

Finally, in Chapter 6, we provide an extended discussion about our findings and try to position them within similar research.

We first provide a summary of our combined results and discuss the insights they offer into the functional 'ingredients' necessary for effectively promoting curiosity-driven behaviors. Specifically, we examine the roles of evaluative and monitoring metacognitive skills, linguistic QA skills, and social perceptions of curiosity and QA. We also critically evaluate the effectiveness of our training methods in fostering these dimensions and their potential to help teachers and EdTech designers nurture students' autonomy and engagement in learning through training curiosity.

Specifically, we question the validity of our methods when used with more formal learning materials and standard educational settings. For instance, can teachers draw inspiration from these findings and adopt our approaches to their formal activities with students? How can we use these findings to motivate the development of specific educational technologies that keep students engaged without relying heavily on features such as extensive gamification? These are some of the critical questions we address to ensure our methods are both practical and adaptable in real-world educational contexts.

We also give perspectives about technical features we could implement to reinforce our approach, specifically giving instant feedback to children about the divergence level and quality of their questions, guesses, etc. This feedback could help children refine and adapt their performance until they achieve an optimal sense of learning and competency. We discuss the first steps we took in this direction by collaborating with researchers in the NLP field to automate the assessment of questions' quality and divergence level using LLMs. By exploring this direction, we study the feasibility of using LLMs to give children instant feedback about their inquiry-based strategies.

Finally, we open discussions about the importance of curiosity-driven QA and metacognition with the current exponential rise of Generative AI (GAI) and the general enthusiasm around its use in education. We wonder how we could adapt the methods presented in this thesis to help children remain curious learners, active and in control of their learning, when interacting with these new complex systems that facilitate access to information. We make hypotheses about the need for curiosity and metacognitive skills to use GAI efficiently during learning and propose future studies to explore this relationship.

Part II

THEORETICAL UNDERPINNINGS

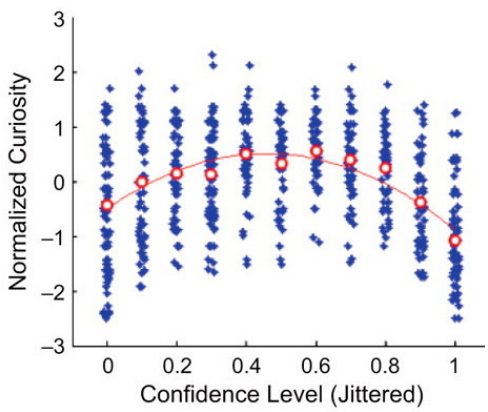
2

CURIOSITY, METACOGNITION AND THEIR PROMOTION IN TODAY'S CLASSROOMS

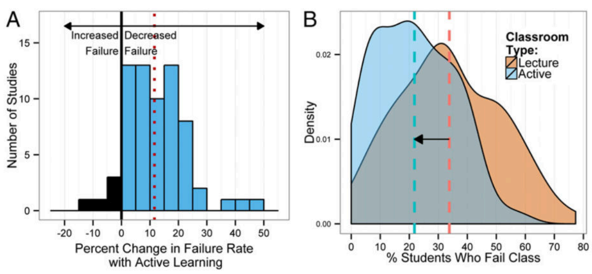
Aims: This chapter aims to explain why we think it is relevant for today's educational system and Educational Technologies industry to focus on promoting curiosity and metacognition in students. It also aims to highlight the potential role of new technologies to fulfill this goal.

Abstract: This chapter starts by providing an overview of curiosity theories, explaining its mechanisms and impact on cognitive development. A key curiosity-driven behavior, question-asking, is highlighted as well, since it is a focus of this thesis. Next, it introduces metacognition, its functions, and its connection to curiosity. Finally, it explores the importance of implementing curiosity-based learning and teaching approaches in the classroom, and discusses how new technologies, particularly GAI, can be used to facilitate such strategies.

Contents



(a) Curiosity is related to metacognitive judgement [117]



(b) Curiosity-based teaching approaches enhance academic achievement [65]

2.1 An overview of information-seeking behaviors and curiosity in biological systems

Information-seeking behaviors are common among humans and non-human primates. They can range from simple actions like passive staring to more complex endeavors such as travelling or pursuing a PhD. Primates devote a significant portion of their lives to seeking new information, a behavior formally known as "exploration" or "orientational-investigatory activity" that can be driven both by extrinsic or intrinsic factors [125]. Notably, information-seeking behaviors can occur even in the absence of direct inherent benefits, such as food or money, indicating that these behaviors are not solely motivated by instrumental or strategic gains.

Broadly speaking, we refer to this form of information-seeking as "curiosity," the intrinsic drive for novel sensations and experiences with the potential for learning [104]. Or, as Kashdan describes it, the "impulse towards a better cognition" [121]. Curiosity is what drives a bird to approach an unknown object, even at the risk of its own safety, to better understand its environment [101]. It motivated explorers like Ibn Battûta to travel over 100,000 kilometers through arid and hostile territories, fully aware of the dangers, to discover the world. Curiosity is a foundational factor that led scientists to make the major discoveries that shape our society today.

With its pervasiveness in every-day life and its critical role for growth and survival, curiosity has sparked the interest of researchers from various fields ranging from psychology to anthropology and computer science. While these efforts have begun to unveil the mechanisms of curiosity, much remains to be fully understood.

More specifically, we still lack a clear and widely accepted definition of curiosity. Indeed, the long history of debates over its nature started with behaviorist approaches that presented curiosity as a basic biological drive that "just is" [221]. These approaches likened it to the pursuit of food or other goods; but instead of these tangible rewards, it drives organisms to pursue non-tangible ones [90]. However, such views lost popularity because they fail to advance discussions on what triggers curiosity when no imminent tangible benefits can be seen/ expected.

Another classic perspective that significantly influenced curiosity research is that of psychologist Daniel Berlyne [24]. With a different approach, he presented two major biological reasons to explain this intrinsic drive. In doing so, he introduced two types of curiosity: 1) specific curiosity, which occurs to solve an uncertainty and/or a conflict facing the organism. It is the direct result of an informational lack detected by the brain during its interactions with the environment. It causes the organism to experience discomfort and motivates it to access novel stimuli that may contain this missing information [90].

And 2) diversive curiosity, which represents the spontaneous pursuit of novel stimuli without needing previous information about it (as opposed to specific curiosity). It is motivated by the need to seek new stimuli, different from the ones presented naturally in the current circumstances (e.g. because these latter only offer too easy or too complex stimuli to process). This need may arise after consistently experiencing monotonous or unchallenging stimuli which can have a negative impact on various psychological functions. Consequently, the organism is driven to seek new stimuli that offer optimal levels of surprise, variety, complexity, and change.

Finally, in more contemporary views on curiosity, we find an emergence of a consensus presenting it as a form of information-seeking behavior that is driven by the expected rewarding value of the information itself [60, 147, 176]. Similar to foraging for food or basic goods, such theories suggest that curious individuals seek novel stimuli for different reasons that can be internal, external, conscious or unconscious.

As some of these explanations remain species-general, Berlyne also proposed another taxonomy to distinguish human curiosity: perceptual vs. epistemic curiosity. The main difference resides in the satiation condition: for perceptual curiosity, the drive to the novel stimuli decreases with the ongoing exposure to it. While for epistemic curiosity, the drive goes beyond resolving an uncertainty or a discomfort, to wanting to acquire new knowledge or a "better cognition" [105]. While perceptual curiosity is species-general, epistemic curiosity is more specific to humans.

2.2 EPISTEMIC CURIOSITY (EC) IN HUMANS

Despite the absence of a unique, agreed-on definition of epistemic curiosity (EC), researchers generally tend to define it as a specific form of intrinsic motivation (IM) that supports and enhances experimentation, exploratory behavior and active learning that is directed towards the acquisition of new knowledge [75, 175]. Unlike motives that lead to seeking information for its instrumental value, this form of IM drives individuals to explore stimuli for their mere nature and "what they are" [51], i.e. their intrinsic value. Epistemic curiosity is considered to be a powerful driving force that builds some of the crucial pillar stones for human civilization and evolution, namely scientific discovery [210] and education [50, 106].

2.2.1 *Main theories of EC*

2.2.1.1 *EC as a product of uncertainty and informational gaps*

Different theories of EC have emerged over the decades. For example, Berlyne presented the conflict theory, in which EC is introduced as a motivational state that drives individuals towards knowledge [23]. According to this theory, EC is satiated by the obtaining of the pursued knowledge (i.e. the learning goal) and is predicted by the initial level of uncertainty surrounding it: the more conflicting a situation is, the more EC can be experienced.

More recently, other theories have emerged such as Lowenstein's "knowledge gap" [71, 147]. In this theory, EC is defined as a cognitive deprivation that arises from the individuals' perception of a difference between a wanted state of knowledge and the current state of knowledge (i.e. a knowledge gap), as evaluated by the individuals themselves. This aligns with the general observation that exploratory behaviors tend to decrease over time: the younger we are (i.e. little we know about the world), the more discrepancies we will detect, and the more information we will want to seek.

However, this trend does not always hold, as older individuals also engage in explorations, but in a more structured manner. This structured exploration in adults can be explained by Lowenstein's suggestion that individuals choose do not engage in exploratory behaviors automatically upon identifying a knowledge gap. Instead, they rely on their subjective perception of the gap's intensity [147]. When a knowledge gap is too large (i.e. the wanted information is completely unknown), they will be demotivated to seek it, same for very small gaps. Therefore, an optimal level of uncertainty is required to stimulate information-search behaviors.

Recognizing such optimal uncertainty experiences requires, among other skills, metacognition that does not until later in life, thus explaining the developmental differences in exploratory behaviors.

Figure 2.1 shows an example of theoretical predictions that explains how feelings about knowledge gaps can impact information-seeking [71].

This theory was supported by a study by Kang et al. [117] showing that individuals are least curious when they have no clues about the answer or when they are extremely confident about it. Such mechanisms are suggested to be the principal motivator for children's curiosity-driven exploratory behavior starting from the age of 4 [164]. Lowenstein's information-gap theory also explains why curiosity is often associated with stimuli that are judged as surprising, novel, violating previous beliefs, etc [18, 176, 182].

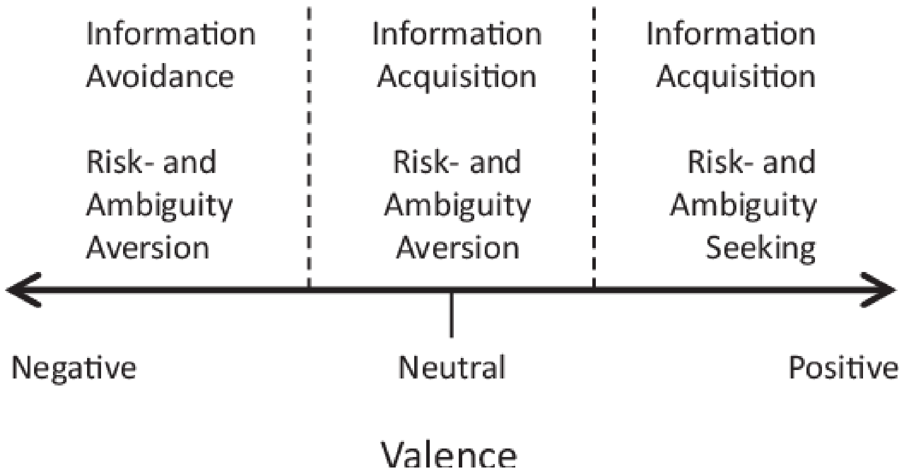


Figure 2.1: Theoretical predictions of informational preferences arising from feelings about information gaps. Golman and Loewenstein, 2018 [71]

2.2.1.2 *EC as a product of the expected learning gains*

In the theories above, EC was presented as directly resulting from the emergence of a particular knowledge gap. In this section, we present an alternative perspective where EC is hypothesized to stem from the anticipation of reward or pleasure associated with acquiring specific information.

Indeed, this idea has led to the development of various explanations of EC's mechanisms. One popular model suggests that individuals predominantly seek tasks of intermediate difficulty level for exploration (not too easy, nor not too difficult). Empirical investigations have supported this idea in various contexts, including infant attention [126, 182], trivia questions [17], etc.

With a more process-like approach, new curiosity theories have also emerged, such as the learning progress (LP) theory [176]. According to this theory, individuals are mostly curious about tasks that would maximize their learning progress: they would make hypotheses about the activities available to them in their environment and choose to pursue the ones that they predict can maximize their learning progress and bring them closer to the mastery of the knowledge component at hand. Here, the motivation is thus more focused on the LP rather than on the final information itself. By applying such algorithms in naturalistic settings, it is suggested that learners can avoid wasting cognitive resources on hazardous tasks that are either already mastered or impossible to master with their current knowledge state [218]. This idea for curiosity-driven exploration has also been applied and shown to be relevant for artificial systems in different domains such as automatic curriculum learning [64], scientific discovery in complex systems [187] and educational applications [44].

In other work, Murayama et al. [168] proposed an operational framework that takes a "process account" of curiosity during the cycles of autonomous learning and effectively "gives up" on the idea that EC can have specific and unique characteristics. This work proposes a framework that captures EC as a process that accompanies sustainable and continuous knowledge acquisition. This framework, greatly inspired from reinforcement-learning models [49] and EC psychology [147] presents autonomous knowledge-seeking as a process initiated by the computation of an expected reward associated with the closing of a recognized knowledge gap. The information resulting from this information-seeking then serves a reward by expanding the individual's knowledge base. It also forms a positive feedback as it facilitates the awareness of further knowledge gaps and thus, the future exploratory behavior (see figure 2.2).

Although not stated explicitly, it is inferred that this model assumes the presence of specific metacognitive skills in individuals (e.g. becoming aware of knowledge gap, monitoring learning progress, etc). Without these, several components of the framework cannot effectively operate. In this PhD, one of our aims will be indeed to make these links more explicit in order to facilitate the design of metacognition-based curiosity training interventions.

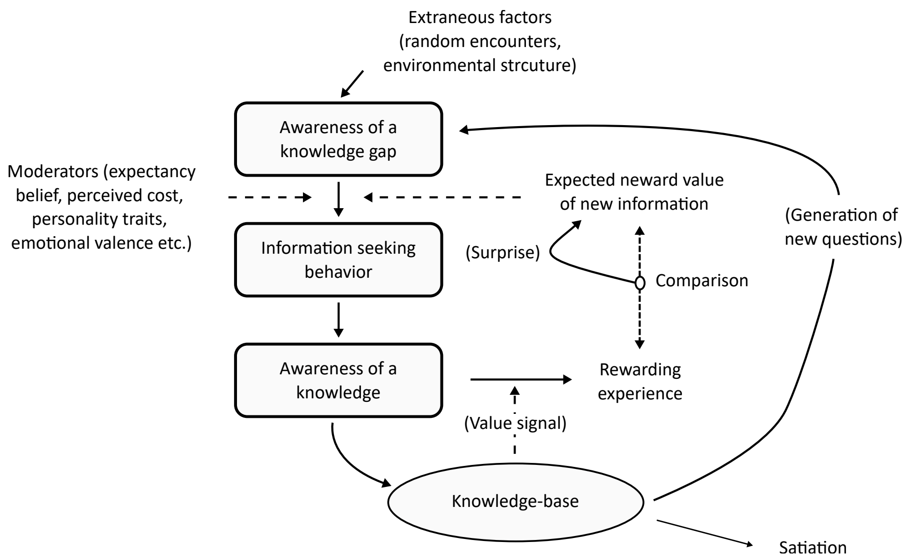


Figure 2.2: Framework for the process account of curiosity in autonomous learning acquisition. Murayama et al., 2019 [168].

Finally, in their rational analysis of curiosity, Dubey and Griffiths [55] present curiosity as a cognitive mechanism that can be responsible for leading people's exploration in a way that maximizes the usefulness of the knowledge. As opposed to the previous theories described above, the authors here propose that the determinants of curiosity-driven explorations can change depending on the environment: people can choose to follow stimuli with high vs. intermediate uncertainty or difficulty depending on the rationality of this decision (see Figure 2.3).

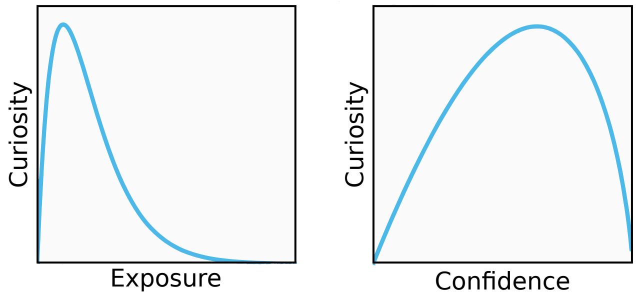


Figure 2.3: Example of a curiosity model showing its links with subjective feelings of knowledge and exposure. Dubey et al., 2020 [55].

2.2.1.3 *Individual differences in EC*

In all models described above, curiosity is represented as a state triggered by specific stimuli in the environment. But it is somewhat inferred that this state also interacts with curiosity trait.

Indeed, individuals' trait curiosity— the differences in their affective responses to uncertainty, challenges and subjective feelings of a lack of expertise— can modulate their information-seeking behaviors when they are faced with the curiosity-eliciting stimuli described in the models above [60]. For example, studies such as in [108] show that children who are more curious by trait are more comfortable with high levels of uncertainty and ask more questions. In another study, children with high curiosity trait reported experiencing greater feelings of presence during exploration [211].

In investigating curiosity as a personal trait, researchers like Litman [141] suggested introducing EC as an individual trait of personality that could have two types: 1) type I (Interest) which is associated with an interest to explore/ discover unfamiliar subjects and enjoy the intrinsic joy that is predicted to result from them. And 2) type D (Deprivation) that is associated with a desire to reduce unwanted states of uncertainty and ignorance. Unlike Type I curiosity, Type D curiosity is oriented to search for a specific unknown, and is only satiated by the availability and accuracy of the response to this unknown.

2.2.2 *Developmental changes in EC and its link to learning*

EC and curiosity-driven learning are present in humans' lives since very early stages and continue to develop throughout childhood and adolescence [59]. Extensive research has focused on studying them in early stages of childhood. They show that young children use the contingencies they experience and their previous knowledge to explore their environments and test relevant hypotheses about unknown objects and reduce uncertainty [75, 83, 125, 200].

For example, infants are more driven to activities with intermediate difficulty rather than easy ones [19, 126, 149, 172] and choose to play with toys that they do not completely understand [205]. Additionally, it has been shown that young children learn better and expand their knowledge more when they are confronted with stimuli that violate their prior knowledge [215], a stimuli property that is suggested to trigger EC as discussed above.

These findings are aligned with the idea that curiosity facilitates learning. Famous child psychologist Piaget for instance considered curiosity and exploration as pillars for cognitive development and key to acquiring knowledge about the world [180]. Moreover, Vygotsky has famously suggested that children's cognitive abilities are not innate or "pre-set", rather, they are developing in proportion to their interactions with the world [225]. Indeed, during these interactions, and with the presence of a "more knowledgeable other", children are encouraged to explore novel stimuli, which can help them adjust and extend their internal models of it and, thus, make learning gains.

Evidence about curiosity and its role in learning for adults has also been an area of research. For example, Wade et al. suggest that learning is best predicted by both curiosity and previous knowledge [226]. Other recent studies in neuroscience also suggest that people are better at learning the information they report to be curious about. In fact, they show an association between these states and the activation and consolidation of long-term memory regions [84].

However, much less is known about the development of curiosity and its interaction with learning in later stages of childhood and adolescence. Indeed, the role of curiosity is very often praised in education [58, 109, 175] but these studies rarely assess curiosity directly. In bridging this gap in literature, Fandakova et al. [59] conducted a study with children and adolescents (from 10 to 14 years old) using a trivia paradigm. Participants had a series of trivia questions that were associated with various curiosity levels. They were first asked to rate their curiosity about these questions, encoded the correct answer and finally, rated their subjective interest in the answer once it was revealed to them. The results showed that participants from different age groups remembered better the answers to the questions they were curious about. More interestingly, they show that the high post-interest scores in the answers reinforced memory beyond the curiosity effects, but only for the adolescents (13-14), not the younger participants (see Figure 2.4). This is in line with research findings suggesting that cognitive processes associated with interest change across development [97].

Finally, evidence concerning developmental changes in curiosity comes also from neuroscience. For instance, in their PACE frame-

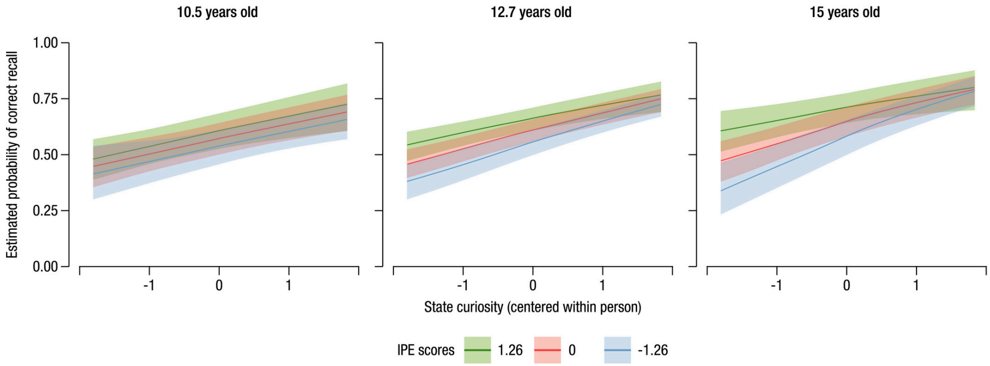


Figure 2.4: Developmental changes in the relationship between curiosity and learning. Fandakova & Gruber, 2021 [59].

work, Gruber et al. suggest that developmental changes in curiosity may be attributed to the maturation status of the lateral prefrontal cortex, a region that is associated also with metacognitive skills such as the ability to self-reflect and regulate learning and memory outcomes [85]. Several studies propose that this region develops in early childhood, begins maturing around the age of 9, and continues to develop throughout adolescence [12, 232].

2.2.3 *Measures of EC*

Due to the numerous conceptualizations of EC, various scales have been developed to assess it, both in general and specific contexts such as education, workspace, et [227].

General domain measures of curiosity include scales such as the "curiosity as a feeling of interest" scale [144]. Highly inspired by Berlyne's theory [24], it addresses curiosity as a personal trait rather than a state and assesses it based on two dimensions: specific and diversive curiosity. Litman also introduced the "Curiosity as Feeling of Deprivation" scale, which accounts for curiosity as an individual difference emerging from discomfort over lacking access to specific information [143]. Drawing from these two studies, Litman also developed the I- (interest) and D-type (deprivation) EC scale [141].

Similarly, we find other scales such as the curiosity and exploration inventory [119]. Also inspired by Berlyne's theory, this scale assess curiosity traits following two dimensions: exploration (i.e. the tendency to seek out new experiences/ information) and absorption (i.e. the tendency to engage in novel explorations). In a more recent study, Kashdan et al. developed the five-dimensional curiosity scale to create a comprehensive multidimensional curiosity trait measure [121]. In this scale, we find dimensions concerning: joyous exploration, thrill-seeking, deprivation sensitivity, stress tolerance and social curiosity. Specific domain curiosity measures were also developed covering a range of domains: science curiosity [228], workplace curiosity [118], social curiosity [189], interpersonal curiosity [133], etc.

On the other hand, several previous studies have also worked on developing behavioral measures of curiosity states in order to avoid the challenges surrounding self-report measures. Example measures include pupils dilatation, exploratory eye movements [135], fMRI imagery[84], ability to ask questions in educational contexts [8, 110], ability to support and manipulate uncecrtainty [111], etc. Furthermore, in analyzing curiosity-driven explorations, studies such as [150] proposed to investigate individuals' patterns and styles of exploration of Wikipedia content after 5 hours of browsing. This study has linked the types of exploration to types of curiosity, i.e. individuals with specific curiosity explored connected content while individuals with higher diversive curiosity explored more diverse content. Finally, authors in [82] proposed to use the willingness to incur a cost in exchange for information as a new behavioral measure for curiosity (e.g. willingness to pay money, wait for information, make an effort in exchange for information, etc).

2.3 QUESTION-ASKING AS A CRUCIAL FORM OF EC

As discussed above, humans develop their abilities to seek information when faced with uncertainty from a very young age, through different behaviors such as staring, pointing, exploring, etc. They begin to vocalize this uncertainty starting at the age of 12-to-24 months [192]. These behaviors are considered to be crucial for children's cognitive development [72, 180].

2.3.1 *Role in cognitive development*

In consistency with the knowledge gap theory, it is suggested that question-asking behaviors can relieve states of curiosity: they are motivated by the desire to make learning progress concerning stimuli that are identified as novel, surprising, etc [25, 79]. They can allow children to address more knowledgeable peers (e.g. parents, caregivers) in an explicit manner in order to help them find answers to the uncertainties they've identified and, therefore, to develop/ extend their understanding of the world [80].

Furthermore, several studies support the idea of question-asking behaviors being a crucial tool that helps individuals shape their information-search and optimize its efficiency for learning. For example, Ruggeri et al. [194] show that children, as young as 10 years old, are able to dynamically adapt the questions they ask, in a way to move closer to the information gains they are expecting, by leveraging the answers they get from external agents. This is an important characteristic of question-asking dynamics as it helps limit the search space that is initially quasi-infinite in the world and requires efficient navigation [113].

Similarly, research such as in[162] suggests that children tend to seek more information when they are faced with explanations that don't contain the answers to their questions, as perceived by the children themselves. Beyond being a testimony for children's question-asking abilities, these findings also suggest that the dynamics of this behavior (ask a question, get an answer, evaluate answer, ask further questions) is a facilitator for an efficient information search directed towards maximizing learning.

2.3.2 *Developmental changes*

Although we know that question-asking skills emerge from early childhood, they are not fully developed until late childhood [113]. Indeed, when looking into children's question quality throughout childhood, researchers in [113] identify three stages: 1) by the age of 5, the ability to recognize "good" questions but not to generate them (here, a "good question" is defined as one that contains the appropriate formulation that can lead to finding the answer to satisfy the informational need expressed by the question [161]). 2) The ability to generate these "good" questions by the age of 7 and, finally, 3) the ability to dynamically adapt their question-asking behaviors to navigate efficiently through the knowledge space and acquire the information needed in an optimal way, by the age of 10.

See Figure 2.5 for an example of the differences in the effectiveness of children's questions between the ages of 4 and 6.

In trying to understand the processes underlying this developmental trajectory, researchers suggested several elements such as metacognition, verbal skills, theory of mind, executive functions, etc [113]. For example, Ronfard et al. [192] link children-under-5's failure to generate precise and goal-directed questions to their lack of executive functions skills that can allow them to coordinate the simultaneous processes needed to do so: identify learning goals (i.e. knowledge gaps), identify reliable information sources (i.e. other agents' knowledge states) and formulate the corresponding inquiry. Furthermore, we also see that metacognitive skills (that develops throughout childhood [78]) are important important in the QA process.

The development of metacognition (i.e. the individual's knowledge about their own learning and ability to regulate it) can indeed facilitate children's capacity to recognize their need for help, adapt their information-search behavior to meet this need by asking the appropriate questions, etc. It is indeed suggested that the development of metacognitive skills helps better understand the features that make the "informational effectiveness" of a question [194].

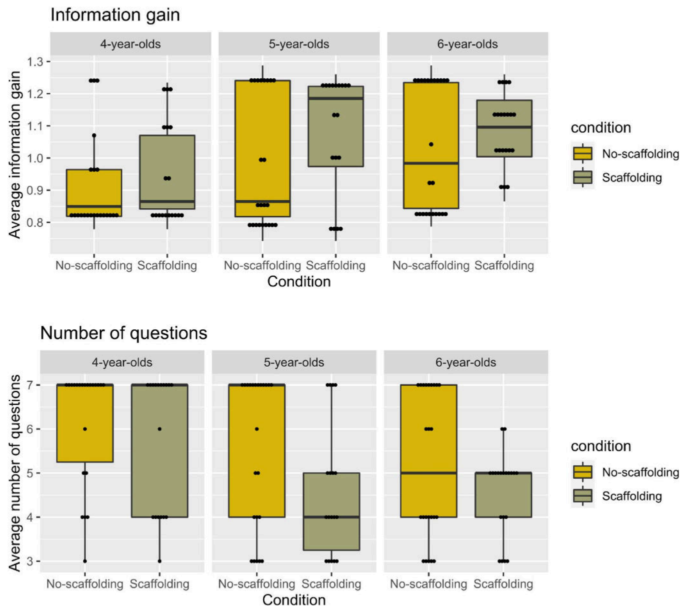


Figure 2.5: Age and scaffolding help children ask more effective questions. Ruggeri et al., 2021 [195].

2.3.3 *Taxonomies of questions and association with curiosity*

Broadly speaking, we often associate question-asking with curiosity in an automatic way, without thinking in detail of the question's purpose, function, etc. But in doing so, do we mean that rhetoric questions can also classify as curiosity-driven behaviors for example? This suggests that defining "good" questions solely as "questions that can eliminate ignorance" is insufficient [123], and more nuanced classifications of questions are necessary.

Lower Level	Cognitive memory questions	*Rote memory; recall of prior learning; recognition of information*
	Convergent thinking questions	*Integrating information; analysis of ideas; synthesising data*
	Divergent thinking questions	*Generating new ideas; putting forward new ideas / views; recognising more than one possibility*
Higher Level	Evaluative thinking questions	*Quality assuring thinking; making judgements; decision making*

Figure 2.6: An example of a taxonomy of questions types proposed by Newton et al., 2017 [171].

To distinguish between questions types, Kearsley et al. presented a non-exhaustive framework that classifies questions with respect to their functions [123]: 1) echoic: used to ask for the repetition of confirmation of an already-encountered; 2) expressive: a question used to convey an attitude (e.g. "aren't you coming?" which usually indicates surprise/ disbelief); 3) social control: questions used to maintain the control during a discussion; and 4) epistemic: questions used to acquire new information. With this framework, one can see that only the latter type of questions (i.e., epistemic) could be related to EC: they allow individuals to seek new information and help them extend their knowledge.

For questions with an epistemic function, researchers have proposed even more thorough distinctions, relying on the relationship between the questions and the answer. For example, Raphael et al. [186] suggested three categories for epistemic questions: questions to which the answer immediately available, questions that require more effort like linking two available information to deduce an answer. And finally, questions that depend on the ability to use the previous knowledge and generate hypotheses. The second and latter categories are likely to be more associated with EC as they require cognitive processes that are associated with it: using previous knowledge to seek further useful and relevant information, making hypotheses, etc.

Similarly, Gallagher and al. [69] proposed two categories: divergent- and convergent-thinking. The first being surface-level questions that require children to explain or compare ideas while the second involves divergent thinking processes and require prediction, making hypotheses or judgements. Alaimi et al. [8] investigated the effect of curiosity on students' questioning abilities and found that the more curious children are by trait, the more divergent-thinking questions they will ask.

See Figure 2.6 for an example of questions classification.

2.4 METACOGNITION: DEFINITIONS, FUNCTIONS AND LINKS WITH CURIOSITY

2.4.1 *Definitions and developmental trajectory*

Metacognition is an important concept across various domains, including biology, psychology, educational cognitive science, and so on. Similar to curiosity, empirical evidence suggests that metacognition positively influences learning and development [213, 224]. However, the reasons and mechanisms underlying this phenomenon remain unclear given the lack of a clear and operational definition of metacognition [134].

In general, metacognition is described as the process of "thinking about thinking" or the "cognition of cognition" [61]. To have a more specific definition, one can examine the meaning of the two parts composing the word "metacognition": the prefix "meta", suggesting an act of "going beyond" cognition. And "cognition", the mental representations and processes that enable an organism to reach its goals in the environment [22]. Combining these definitions, metacognition can be defined as the process and resulting products through which individuals evaluate and monitor their own knowledge states and performances, enabling them to regulate their future cognitive activities and behavior [170].

More particularly, according to Flavell, metacognition has two main components: metacognitive knowledge (declarative metacognition) and metacognitive experiences (procedural metacognition) [61]. 1) Metacognition knowledge refers to the individual's knowledge about their own cognitive tasks, goals and strategies. An example would be an individual recognizing they are better at learning mathematics than arts for instance [88]. 2) The procedural component refers to conscious experiences that occur during cognitive activity. For example, the feelings that we don't understand something when someone is talking, feelings of knowledge, etc. This component can activate strategies of self-regulation and active monitoring and control of performance in a task [204].

See Figure 2.7 for a proposition of metacognition components, presented by Schneider [204].

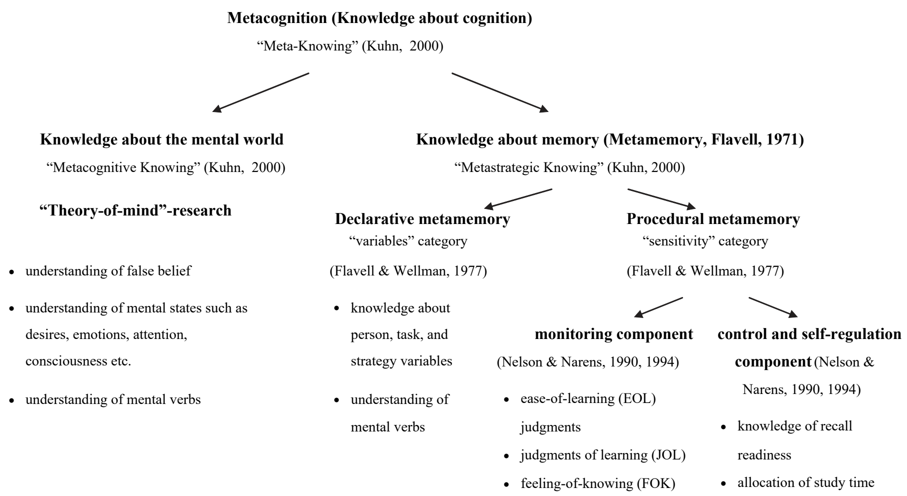


Figure 2.7: A taxonomy of metacognition components proposed by Schneider [204].

Based on these definitions, we often tend to associate metacognition with situations such as people discussing their thoughts about their performances in a certain task, speculating on their potential success, etc. This has led us to perceive metacognition as a declarative skill primarily manifested in behaviors like self-reflection and self-reporting. And since self-reflection is considered as a rather advanced behavior [30], researchers have long assumed that metacognition is only available to human adults [39].

However, this view has been challenged by other researchers who have suggested the existence of more fundamental or "core" forms of metacognition observable from the early stages of human life [76].

For instance, Kouider and Goupil [78] demonstrated in a study that even pre-verbal infants (aged between 18 to 20 months) can report and monitor their own uncertainty using pointing gestures. Other research showed that babies use pointing in a interrogative manner [19], or staring to communicate uncertainty [31], etc. Taken together, these findings suggest that individuals, from a young age, use core forms of metacognition to adapt their learning strategies to their objectives.

In Shea et al. [206], the authors proposed to explain this by introducing a 2-system framework for metacognition: 1) an "intra-personal" system that operates implicitly within the individual to guide their processes. This system is shared with human adults, infants and even some non-human animals. 2) a "supra-natural" system that enables individuals to share their metacognitive representations (see Figure 2.8). This latter is unique to humans, develops in later phases of human life (not before 3 years old according to research in [31]), and is also known as "explicit self-reflection" [76].

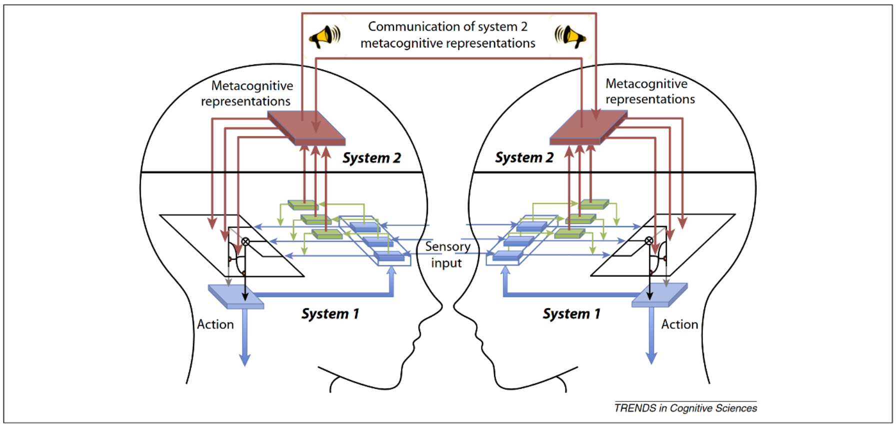


Figure 2.8: Sharing of metacognitive representations between two agents: system 2 MC representations are in verbal form and are derived from system 1 MC where the former are used to improve control. Shea et al. 2014, [206].

2.4.2 *Links with epistemic curiosity*

As discussed earlier on, curious information-seeking behaviors remain somewhat complex, especially regarding their triggering conditions. In trying to explain them, Lauriola et al. leveraged our understanding of metacognition and proposed the idea that curious exploration is, at its core, a metacognitive self-regulatory process [136]. Goupil et al. also second such a view on curiosity and defend the idea of "curiosity as a metacognitive feeling" [77]. They emphasize two inherently metacognitive processes as the triggering elements for curiosity: 1) the assessment of one's informational needs, and 2) a prediction of the learning gains they can achieve by pursuing some specific information. See an example of a framework presenting the link between curiosity and metacognition in Figure 2.9.

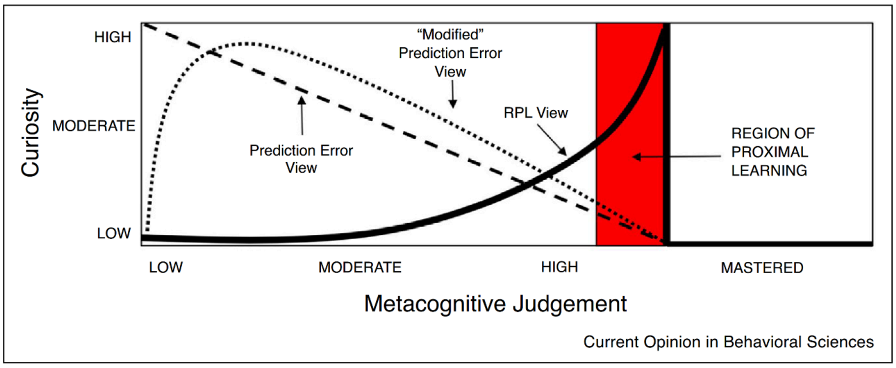


Figure 2.9: The link between curiosity and metacognition within the Region of Proximal Learning framework, proposed by Metcalfe et al., 2020 [159].

This explanation aligns with the curiosity theories and frameworks discussed in the previous chapter. For instance, if we consider Loewenstein's Information Gap Theory [147], we can see that its basis is metacognitive. Indeed, according to this theory, individuals must evaluate their own knowledge within their immediate environment, identify a knowledge gap by recognizing an un-mastered competence or incomplete information, and subsequently decide to seek stimuli that may address this identified knowledge gap. These three steps represent early, "core" forms of metacognition [31] that begin to develop in humans since their early years (around 18 months). They can explain the curious exploratory behaviors in infants since a very early age.

In the same line, the learning progress theory proposed by Oudeyer et al. [176], suggests that individuals determine which information to seek based on their assessment of how this information will optimize their learning progress. This implies that learners regulate their curious information-seeking behaviors through metacognitive monitoring and regulation procedures: first, by assessing their current knowledge state, and second, by predicting the potential learning outcomes that could be achieved by pursuing specific information or stimuli. Similarly, Murayama's process account of curiosity [168] in autonomous knowledge acquisition suggests that curiosity-driven information-seeking is primarily internally modulated by an expected feeling of reward that individuals predict to obtain from their future inquiries. Generating expected rewards, such as making educated guesses about unknown information, can also be viewed as a metacognitive procedure, as it involves integrating previous knowledge to predict how best to enhance future cognitive outcomes.

These links between curiosity and advanced metacognitive skills have also been suggested through neuro-cognitive models such as PACE [85]. This work suggests that curiosity is elicited based on the developmental stage of the lateral PFC, which is directly linked to the metacognitive abilities to self-reflect and to self-regulate learning [59].

Taken together, these different views suggest that curiosity-driven information-seeking behaviors can depend on metacognitive processes such as the individuals' ability to become sensitive to their informational needs and monitor the progress of these latter. But while extensive research supports this idea, other other perspectives propose that curiosity can be seen as a simple motivational experience that does not necessarily require the ability to make meta-representations [39].

2.5 CURIOSITY IN THE CLASSROOM

Since we took the first steps towards understanding curiosity and its impact on learning, the need to promote it in educational settings has been increasing [109].

For instance, Friedman has suggested that "curiosity combined with motivation to learn are more important than intelligence" in schools [66]. While this assertion may be considered somewhat oversimplified, it is widely acknowledged that curiosity and the motivation to learn play a crucial role in shaping learning outcomes and experiences. Indeed, several lines of research have now linked enhanced academic performance to curiosity indicators such as active engagement and interest [203], question-asking behaviors [111], as well as to parental encouragement and nurturing of children's curiosity [74] (see Figure 2.10).

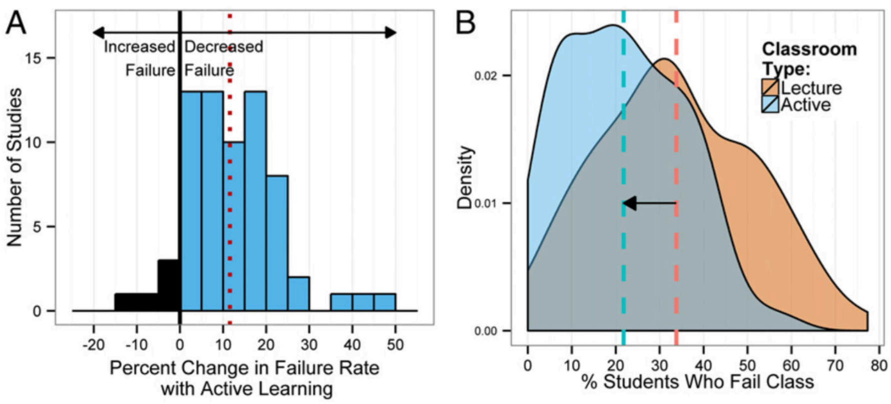


Figure 2.10: The failure rate in the same course, under active learning directed by curiosity is lower compared to a course directed by standard instructions (average difference of 12%). The same applies to density graphs. Freeman et al., 2014 [65].

Additionally, curiosity has been associated with a state of general well-being and adaptive social behavior by enhancing the sense of self-efficacy and novelty thriving [119, 120]. Meaning that individuals experience greater enjoyment from the learning situation, which drives them to seek even more information and thereby further optimize their learning opportunities. This positive feedback loop is the crux of curiosity's role in optimizing learning, and it is inherently metacognitive: it involves evaluating one's own knowledge and using this assessment to regulate subsequent learning opportunities. Furthermore, metacognition plays a central role in empowering students by granting them agency and enabling them to take control of their learning [152], two concepts that, in turn, are suggested to enhance the benefits of curiosity in learning [8].

2.5.1 *Supporting curiosity in the classroom: inspiration for alternative educational practices*

As theoretical evidence for the role of curiosity in learning is piling up, educational researchers and practitioners started thinking about novel approaches to nurture this skill in classrooms [50]. But similar to the definition of curiosity, the ways to promote it also lack a unique and clear framework.

2.5.1.1 *By promoting inquiry-based approaches*

However, a consensus has been reached concerning the general guidelines to apply in the classroom. One idea is to assist students in regulating their learning and guide them towards recognizing the relevant challenges that may help them maximize learning, by explicitly helping them identify information gaps [58, 111]. This could be implemented in the classroom by shifting from direct instruction-driven teaching approaches to inquiry-based ones. For instance, for an activity around prime numbers, teachers can either ask "Where are the prime numbers in this list?" or say "I wonder, why is it important to qualify some numbers as primes" [109]. Beyond merely retrieving the factual information the teacher asks for, the latter approach can encourage students to generate novel questions that lead to new information beyond what is typically asked of them. These will be questions that are within their competencies but that require external incentives and support to be formulated effectively [58]. The inspiration for such approaches is rooted in Vygotsky's theory about the role of caregivers in children's cognitive development through encouraging exploration [225].

Several other ways could also be implemented to replace standard direct instructions in the classroom. For example, researchers in [36] proposed a novel routine where children learn through exploration: they first formulate a critical question related to the learning material they wish to study. Next, they explore the related available pedagogical resources and formulate subsequent questions about them. Finally, they use these questions and resources to formulate a concise answer to their initial critical question. Results indicated significant differences in learning progress between students who followed this routine and those who underwent traditional instruction-based sessions.

2.5.1.2 *By promoting comfort with question-asking and uncertainty*

Another relevant idea put forward by research is the need to promote "comfort with uncertainty" among students [111]. This means that students should learn how to identify uncertainty and associate it with learning opportunities rather than with stressful feelings such as the fear of failure, which is often the case for young learners [157].

Teachers could adopt different strategies to achieve this. For example, it is suggested that teachers expressing and communicating their own uncertainties, acting on them by exploring information, asking questions and making mistakes, can result in students "catching" and adopting this behavior [81]. This is in line with the learning theories suggesting that children track the knowledgeability of other agents of their environment and use this signal to shape their own information-search behaviors [94, 174, 199].

In other words, if students see their teachers (i.e. the agents they consider as more knowledgeable) expressing uncertainty and acting on it, it is possible that they will do so too. It goes without saying that while feeling comfortable with uncertainty is a first powerful step towards supporting curiosity, it is equally important to instruct students, without imposing specific methods, on how to choose the relevant behaviors that can help them reduce these uncertainties [109]. For instance, researchers have often highlighted the importance of teaching students how to formulate efficient questions to satiate their informational needs [201].

See Figure 2.11 for an example of conditions to enhance curiosity in classrooms, proposed by Peterson et al. [179].

2.5.1.3 *By promoting individualized learning*

Several research has highlighted the need for creating learning sequences that present students with the optimal levels of uncertainty, complexity, novelty, etc that could lead them towards exploration [102]. For such approaches to be efficient, teachers are required to push students to constantly reflect on what they do and do not know and make connections between the two, in order to be able to recognize uncertainty and be motivated to pursue it. This connection puts forward the importance of training metacognition to support curiosity in education: individuals need to be able to self-assess their knowledge states in order to feel states of curiosity and be motivated to search for information. They also need to be able to self-regulate their learning strategies in order to be able to achieve the learning progress needed to relieve these states of curiosity.

	Intervention/condition
Biopsychosocial Factors	Having some, but not too much knowledge
	Believing that knowledge evolves, and requires justification and evaluation
	Growth mindset
Proximal Processes	Opportunities for invention, exploration, and question-asking
	Direct instruction on question-asking and inquiry techniques
	Culturally relevant curricula and pedagogy
Educational Context	Educational settings that decrease emphasis on high-stakes testing
	Classrooms that value positive peer discussions

Figure 2.11: Interventions and conditions associated with increased or sustained curiosity levels in the classroom. Peterson, 2020 [179].

2.5.1.4 *By promoting autonomy and self-regulation*

On the other hand, theories based on leveraging people's intrinsic motivation in order to support their growth and learning have also appeared (e.g. the self-determination theory (SDT) [197]). Such theories had strong implications in the educational field as they were associated with major impact on students' need for autonomy, competence and relatedness [198].

While these theories have inspired the educational systems to propose more pedagogical activities based on inquiry, more 'radical' forms of these approaches have also appeared. A prominent example is the Montessori system, which grants students significant freedom to choose their learning materials and eliminates external rewards such as grades. Instead, it emphasizes the formal training of sensory, motor, and mental capacities to facilitate self-directed learning [165].

Although evidence supporting the effectiveness of this method has been limited, there is a general consensus that children can derive cognitive and social benefits from it. Moreover, studies such as in [139] have demonstrated that 12-year-old students in Montessori schools perform similarly in writing tasks (essays) compared to students in traditional schools while being rated significantly more creative.

2.5.2 *Main challenges*

Despite recognizing the advantages of curiosity and metacognition in learning, teachers encounter significant challenges in implementing educational practices that support these skills. The primary obstacle lies in adhering to overarching educational policies. Teachers are indeed required to have students practice and solve learning problems using uniform and standardized procedures dictated by the educational system they operate within. This lack of flexibility and limited opportunities for students to explore material that interests them can hinder curiosity. Additionally, the use of common and standard assessment measures to evaluate students' performance, which often only assesses their response to specific factual knowledge, can discourage individuals from being proactive in the classroom during the learning process.

Even in the absence of such rigidity, a significant challenge remains in the implementation of curiosity-supporting activities in classroom settings. As discussed earlier, fostering curiosity entails presenting puzzling activities and problems that offer optimal levels of novelty and complexity to stimulate children's curiosity and motivation to seek further information. However, determining this optimal level is contingent upon various factors such as prior knowledge, interests, and individual differences, which greatly vary among students in the classroom. Consequently, to offer activities that cultivate their students' individual curiosity, teachers must develop personalized learning sequences. Yet, fulfilling such requirements is nearly impossible for teachers, given the substantial resources they demand in terms of time, the creation of pedagogical content, and their other responsibilities.

2.6 LEVERAGING NEW TECHNOLOGIES TO BUILD EFFICIENT EDUCATIONAL APPLICATIONS

Over the past decades, technology has transformed many aspects of daily life, including educational practices. Leveraging technology to reinforce the quality of the educational system seems both a highly important and promising track. Indeed, although education presents high stakes for the future and development of societies, it is still lacking the means and resources to maximize its efficiency [89].

By the 70's, B.F. Skinner was already experimenting and talking about the need to have "programmed instructional materials" designed to help students progress and learn at their own pace as an important mean to face the motivational problems seen in schools [212]. In the same context, the educational system has underwent a big change with the arrival of the PLATO system [93]: a generalized computer-assisted instruction system that incorporated several online concepts to help enrich the learning experience and foster collaboration with tools such as forums, online testing, remote screen sharing, etc. The system was used worldwide and was highly successful in inducing more motivation and engagement [93].

These positive results have opened the path for the educational technology sector to grow exponentially, both in research and industrial domains. Indeed, over the past decades, several different tools have emerged to computerize learning sequences and instructions [131], test students' knowledge and provide them with direct feedback [52], etc. More recent forms of such systems that encompass several of these features simultaneously have also appeared (e.g. MOOCs) and proved to be also efficient. They even grew to be essential to meet the needs of education globalization and worldwide crises such as the recent COVID-19 pandemic [89].

See Figure 2.12 for predictions of the most influential educational technologies from 2011 to 2021, as analyzed in the Horizons Reports in 2021.

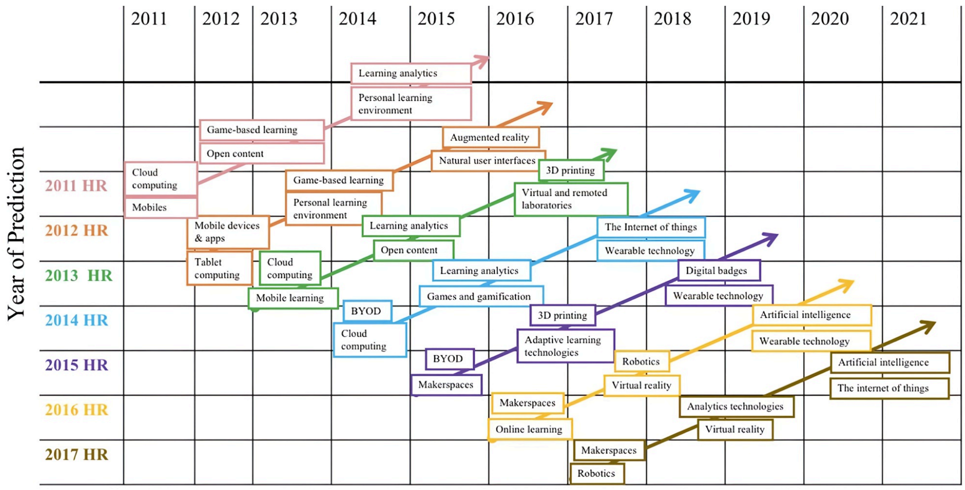


Figure 2.12: Predicted impact of educational technologies from 2011 to 2021, as perceived in 2011. Dube et al., 2022 [54].

Furthermore, and despite several studies demonstrating promising positive impacts of integrating digital tools in education [98], educational teams continue to face a significant issue: the lack of student engagement and motivation. A considerable part of this problem is attributed to the adoption of one-size-fits-all methods typically seen in the classroom, even with the use of new technology [114].

The need for personalized learning experiences and for activities that could be meaningful and of interest for each student has thus been expanding. Beginning with simple experiments, Cordova et al. [47] showed that leaving students the ability to choose their favourite subjects and exercises to work on improved their intrinsic motivation, engagement and learning outcomes in mathematics. Similar results have also been seen in geometry [56]. From there, more advanced systems have surfaced to provide learning sequences that are tailored to every learner's needs and goals: the Intelligent Tutoring Systems (ITS).

A variety of ITSs were developed over the recent years and were applied in different domains: geometry [191], language learning [153], programming [116], etc. While sophisticated, these systems presented several limitations as they require very complex methods such as learner modelling. In this context, research has explored more automatic ways of personalization without the need for particular assumptions about the student. For example, the KidLearn project [46] proposed a system that provides adaptive personalized learning sequences 'online'. Meaning that the trajectories proposed to users are based on their current performance and are dynamically adapted to propose activities that maximize the learning progress. Their results showed indeed greater learning progress compared with 'expert sequences' that are predefined and common for all users.

Furthermore, the system was also tested in settings that enhance students' agency, by giving them choice over the exercises that are being proposed to them. Interestingly, their results showed a positive impact of the students' choice on their intrinsic motivation and learning, but only in the case where curriculum personalization was already effective [45].

Finally, educational technologies have also experienced exponential development with the recent massive advances in the Natural Language Processing (NLP) field. Indeed, work with NLP techniques has allowed the development of new tools to automate the semantic, syntactic and grammatical analysis of students' typed and spoken inputs, the detection of their potential errors, the generation of personalized feedback, etc [37, 140, 207].

These new techniques have also been used to automate several tasks in order to alleviate teachers' workload like generating content for some learning sequences. Authors in [217], for instance, showed that NLP-based methods can be used to extract the key scientific ideas and concepts from science educational resources that, according to human experts, are important for a better and deeper understanding of the domain. Another important application was the generation of quiz content to evaluate students' knowledge [33, 216].

Fast forward to the recent few years, educational technologies have also benefited from the big rise of Generative Artificial Intelligence (GAI). Several new possibilities are now open with the introduction of pre-trained Large Language Models (LLMs) as they demonstrate impressive powers in generating natural language and are easy to use in downstream tasks by different actors, without the need for expertise in AI [34, 53].

Indeed, in a context where students have more diverse needs, and educational institutions suffer from teachers shortages, GAI could play a crucial role. For example, large language models (LLMs) can help educators create pedagogical content with personalized learning sequences and feedback, attractive and collaborative interactions [122], etc. LLMs can also be used to provide students with personalized pedagogical guidance through informative feedback, metacognitive scaffolding, etc [27, 233]. In a different direction, LLMs have also been used to train beginner teachers interacting with students and be more at ease in their function [156]. In this study, authors used ChatGPT to simulate students with different personalities and pedagogical goals and had them interact with new teaching assistants (TA) by asking questions, expressing confusion, low self-confidence, etc. TAs then had the opportunity to answer the students' questions and help them achieve their goals. The TAs that participated in the study reported indeed a fruitful experience and better confidence to teach.

Overall, various forms of technology have progressively made their way to the classrooms to help alleviate the lack of material and time resources that teachers suffer from. But while being considered as a part of the solution for educational inequalities, it can also represent a social barrier between those who have access to these technologies and those who do not [127].

2.6.1 *Focus on new technologies to foster curiosity and MC*

Several studies have introduced the idea of curiosity as a malleable skill that can be trained in children through the promotion of specific behaviors [111]. Moreover, research has shown that the social environment has a paramount influence on children's exploratory behaviors and motivation to learn [58]. From here, the idea of simulating controlled curiosity-fostering environments with artificial agents has surfaced. Such environments can also help children manifest their curiosity better as they bypass social constraints such as the fear of classmates or teachers judgment when asking questions, fear of failure, etc.

In this context, several studies have investigated the efficiency of such methods. For example, Ceha et al. [40] introduced a social robot that acts and expresses curiosity about science topics and studies whether children can 'catch' this curiosity through their interactions with the robot (see Figure 2.13). Their results showed indeed that children were able to recognize the robot's curiosity-driven behaviors and imitate them in their own learning. Similarly, Gordon et al. [73] had children interact with an autonomous virtual agent that exhibited curiosity-driven behaviors. Their results showed that the interaction selectively increased children's aspects of curiosity behavior.

In the same line of work, Alaimi et al. [8] proposed a simulated pedagogical agent that trains children to ask convergent and divergent-thinking questions about a text by giving them specific linguistic and semantic cues. Their results showed that the divergent-thinking oriented support helped children increase their number of higher-level questions and that this performance predicted their trait curiosity as reported by the children's parents. While very promising, these different studies did not lead to significant differences in learning progress, leaving the question of 'whether curiosity can lead to learning gains' in educational settings still unanswered.

As a follow-up to this study, one of the objectives of this thesis will be to investigate the design of artificial agents to train children to ask divergent-thinking questions, use these questions to explore educational resources independently and make learning progress [3]. This study is indeed the first empirical contribution of this thesis and will be presented in details in Chapter 5.

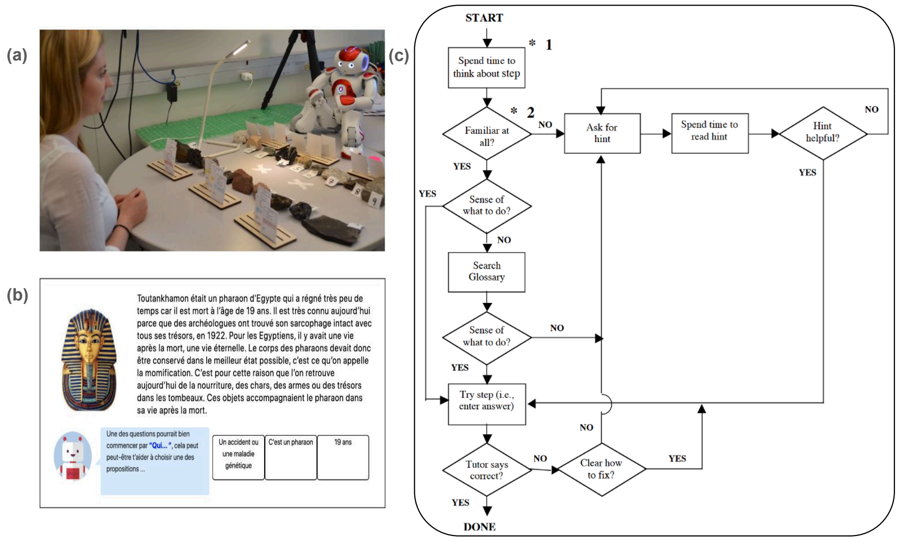


Figure 2.13: Examples of educational technologies to enhance curiosity and/or metacognition. (a) training curiosity by imitation with social robots (Ceha et al., 2019 [40]). (b) training curiosity by practicing question-asking (Alaimi et al., 2020 [8]). (c) training metacognitive monitoring skills by tutoring help-seeking (Aleven et al., 2004 [11]).

A great line of work has also investigated leveraging new technologies to enhance metacognition, an essential component for curiosity, as seen in the previous chapters. For instance, authors in [208] showed that using conversational pedagogical agents could impact children's perception of their own competencies and their self-efficacy. Aleven et al. [9] investigated training metacognitive strategies through an online "learning by explaining" method with an artificial tutor. The computer-based system was implemented to support students in self-explanation, i.e. it encouraged them to explain how they're solving a problem, step by step. Their results showed indeed that students who had the intelligent tutor were more able to explain their steps and were more successful in learning.

Similarly, Azevedo [13, 14] highlighted in a series of work the importance and efficiency of using computer assistance to prompt and support students' self-regulatory learning strategies. In a similar context focused on curiosity-driven learning, one empirical contribution we aim to make in this thesis is to design digital learning environments to train specific metacognitive skills that are directly linked to curiosity. This study is the third empirical contribution of this thesis and will be presented in details in Chapter 5.

2.6.2 *Generative AI for supporting active learning*

Generative AI has been the central theme of discussions in the educational field for the last few months, especially since the release of the popular model ChatGPT. Heated debates about the implications of introducing LLMs in education, the opportunities and risks are animating the research community, the policymakers and the educational corpus.

Indeed, on the one hand, GAI may provide personalized and interactive pedagogical content that could favor students' intrinsic motivation and active engagement [122]. This can be done by designing specific interactions to trigger deeper and higher-level thinking capabilities. Some previous work has indeed shown some promising results concerning such applications. In this context, one of the aims of this thesis will be to evaluate the efficiency of using LLMs to prompt students' curiosity during learning tasks. This study is the second empirical contribution of this thesis and will be presented in details in Chapter 4.

LLMs can also help reinforce the adaptive learning strategies by giving personalized feedback that covers both cognitive and metacognitive performances [27, 233]. Implementing such behaviors could be of a great benefit to students, both on the cognitive and metacognitive levels, as it could help them have the relevant information to make accurate mental models of the pedagogical concepts at hand and to auto-regulate their learning and be more efficient in their strategies to reach their goals.

A new study by Habib et al. [87] used the Alternative Uses Test (AUT) with undergraduate students to investigate the impact of having access to ChatGPT on their performance in this creativity test.
By measuring flexibility, originality, fluency and elaboration both when students had access to ChatGPT to do the test and when they didn't, authors showed rather positive results in favor of using the LLM as a brainstorming partner or a collaborator to support creativity. Authors also showed that while some of the results are positive, there are also negative impacts on creativity and creative confidence, calling for a careful use of this technology in education.

With similar intents, authors in [48] investgated the effect of GAI on students' agency. While they support the idea that GAI facilitates scaffolding personalized learning, they show that students tend to rely on AI rather than learning from it, thus creating a risk for passiveness and loss of control and agency. Furthermore, Macina et al. [151] highlight in their work the fact that, even when giving accurate answers to mathematics problems, the solutions are often given directly and quickly by the LLM, which doesn't give the opportunity for students to think, adapt, etc.

So while GAI is undeniably highly promising for building strong educational applications, several concerns are also raised. Mainly because: 1) GAI challenges the individuals' ability to work for information by leading them to believe that solving any type of problem can be easily accessible without the need for any specific previous knowledge. This is a crucial property in LLMs that can reduce curiosity during learning. And 2) GAI challenges the development of critical thinking and self-reflective behaviors. This risk comes from another LLMs' property that could lead to the loss of learning control: the lack of uncertainty signalling and the continuous confidence they exhibit even in cases of failure. This can lead to an over-dimensioned representation of the knowledgeability of LLMs and thus, to systematically accepting their behaviors without much analytic or critical thinking. It can also lead to over-estimating one's own knowledge [124]. As discussed before, both these behaviors can suppress curiosity and exploratory behaviors given the absence of an uncertainty signal [147].

Taken together, these findings all support the promising path of using LLMs to build educational applications but highlight the need for their alignment with pedagogical goals and policies to maximize their efficiency and reduce the associated risks [4].

Part III

EMPIRICAL CONTRIBUTIONS: DESIGNING ARTIFICIAL AGENTS TO FOSTER CURIOSITY AND METACOGNITION

3 STUDY 1. DESIGNING, IMPLEMENTING AND EVALUATING "KIDSASK": AN ONLINE PLATFORM FOR TRAINING DIVERGENT QA SKILLS

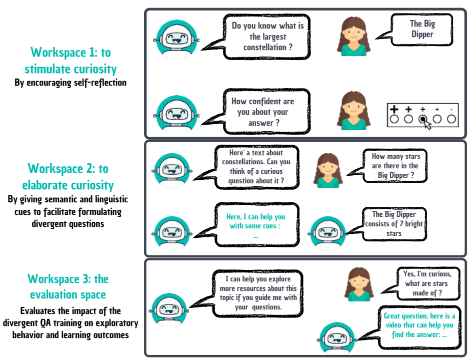
Collaborators: Edith Law, HCI Lab, University of Waterloo, Canada.

Aims: This chapter aims to introduce the design, implementation and experimental methods to evaluate "KidsAsk", an online platform that trains children's divergent QA skills.

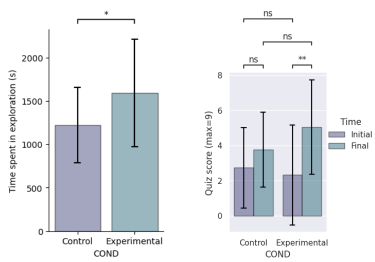
Abstract: This chapter outlines the rationale behind the design choices of "KidsAsk," as well as the methods for its technical implementation. It then details the experimental setup used to evaluate the platform: 51 students aged between 9 and 10 years old used "KidsAsk" for a reading-comprehension task. "KidsAsk" offered interactions with either an incentive agent that encourages divergent QA about the text by offering specific cues or with a neutral one that only checks if children have questions. Finally, the chapter presents the results: the incentive agent significantly enhanced children's divergent QA skills. This enhancement also resulted in longer autonomous exploratory behaviors and stronger domain-knowledge learning progress.

Contents

The work presented in this chapter is published in IJHCS, 2022: Rania Abdelghani et al. "Conversational agents for fostering curiosity-driven learning in children." In: International Journal of Human-Computer Studies 167 (2022), p. 102887



(a) Illustration of "KidsAsk"



(b) Impact on exploration & learning progress

3.1 INTRODUCTION

Curiosity is considered as a crucial driver for learning and cognitive development. In education, it is often seen as a principal factor to keep students motivated and engaged — two factors that can make the learning experience both efficient and enjoyable. One of the main expressions of epistemic curiosity in the classroom is asking divergent questions. These questions, as opposed to memory-based questions, aim to search for novel/ unknown information and are formulated based on deep cognitive reasoning and self-reflection (see Section 2.3 for more details about questions taxonomy).

However, contrary to the popular belief that children are always curious and asking questions in the classroom, previous studies show different observations: students don't ask questions very often and when they do, they ask low-level and memory-based questions (e.g. ask the teacher to repeat an explanation, etc) [80, 103]. Graesser et al. explain this lack of curiosity-driven QA behaviors in children with three main reasons: 1) children do not ask questions because they associate this behavior with negative feelings like stress or anxiety. Indeed, students' encounters with QA behaviors in the classroom happen mostly when teachers are asking evaluation questions. Students' role is thus to answer questions accurately, and not ask them. Another social brake that keeps children from asking questions is their fear of negative judgement from their peers or teachers. It is in fact common to see asking questions as a sign of stubbornness, stupidity, etc [183]. 2) Asking high-level and divergent questions requires rather advanced linguistic skills that children in primary school might still be missing. In their investigation of the questions' quality generated by 5th grade students, Graesser et al. found that the majority of these questions are Yes/No questions, start with basic questioning words (e.g. what, where, etc). However, curiosity-driven and divergent questions are usually associated with more complex starters such as "What is the difference", "How do you link", etc that children are not necessarily comfortable with using [42, 69]. And Finally, 3) children may also refrain from asking questions because they are simply not curious about learning and/or struggle to identify uncertainties in their own knowledge or learning goals they want to pursue [80]. Indeed, several studies suggest that young learners tend to overestimate their skills and, therefore, often miss new opportunities of learning by asking questions [129, 167].

In this context, the first study of this thesis aims to train divergent QA skills in children, by trying to work around the three main problems stated above. We aim to design environments that reduce social anxiety about asking questions (digital environments), offer linguistic training targeted for divergent QA and finally, help children see their knowledge gaps more explicitly. To do so, we propose to design a web-based platform we call "KidsAsk". The platform offers interactions with conversational agents (CAs) to provide children with the support needed to generate divergent questions during reading-comprehension tasks. We choose the CAs technology as it has long been long used in educational settings and showed positive impact on students' engagement and on reducing performance-related social anxiety [10]. Previous work has also suggested that these systems can be a promising path for enhancing curiosity-driven behaviors [8], but little is known about how this could affect learning and domain-knowledge learning progress.

More concretely, "KidsAsk" offers three main activities in three different work-spaces (WS): 1) WS1: designed to train explicit self-reflection in order to help elicit curiosity during learning. This is done by asking children to report confidence levels in their performance during the tasks they are given. 2) WS2 is designed to train the elaboration of curious thinking by helping children transform knowledge gaps into divergent questions. This is done with the help of CAs that explicitly drive students' attention towards potential knowledge gaps (KGs) about the learning material and then offer them linguistic support to facilitate formulating the corresponding question that can compensate for them. 3) WS3 is designed to investigate the impact of the divergent QA training proposed in WS2 on students' ability to explore and navigate through the learning space efficiently and autonomously. We also use to evaluate the impact of curiosity-driven behaviors on the domain-specific learning progress made by participants.

This work contributes both to the ongoing research on novel learning approaches in general and to our understanding of how to design effective interactive platforms for promoting inquiry-based and curiosity-driven learning.

3.2 Design and implementation of "KidsAsk"

3.2.1 *Design of the "KidsAsk" platform*

As illustrated in Figure 3.1, we design the "KidsAsk" platform's training mechanisms following the components of an operational model of curiosity inspired from Loewenstein's knowledge gap theory (see Section 2.2 for details). In this theory, curiosity-driven information-seeking behaviors are suggested to be primarily stimulated by the recognition of a gap in one's own knowledge. This awareness is, in turn, the result of a self-reflection process when the individual encounters a novel/unknown stimuli to learn.

With this in mind, we design "KidsAsk" with three main work-spaces: 1) one to elicit curiosity by encouraging self-reflection. And 2) one to elaborate curiosity by giving the support needed to transform a knowledge gap into a divergent question that can help seek the missing information efficiently. Based on previous research investigating brakes that keep children from being curious in the classroom [80], we choose to provide this help both on the semantic and linguistic levels. Finally, another aim of our work is to study the impact of training curious QA skills on shaping the efficiency of children's autonomous exploration and information-search behaviors, as well as on their learning outcomes. For this reason, we add the third workspace "the evaluation space". 3) a workspace where children can autonomously explore educational content about a specific topic and can ask divergent questions if they want to, without any help from external agents. This space is then used to assess the transfer of the divergent QA skills trained in WS2 in a new learning context, the exploratory behaviors (in terms of length and organisation) and the domain-knowledge learning progress. We also investigate the relationship between these three indicators of learning.

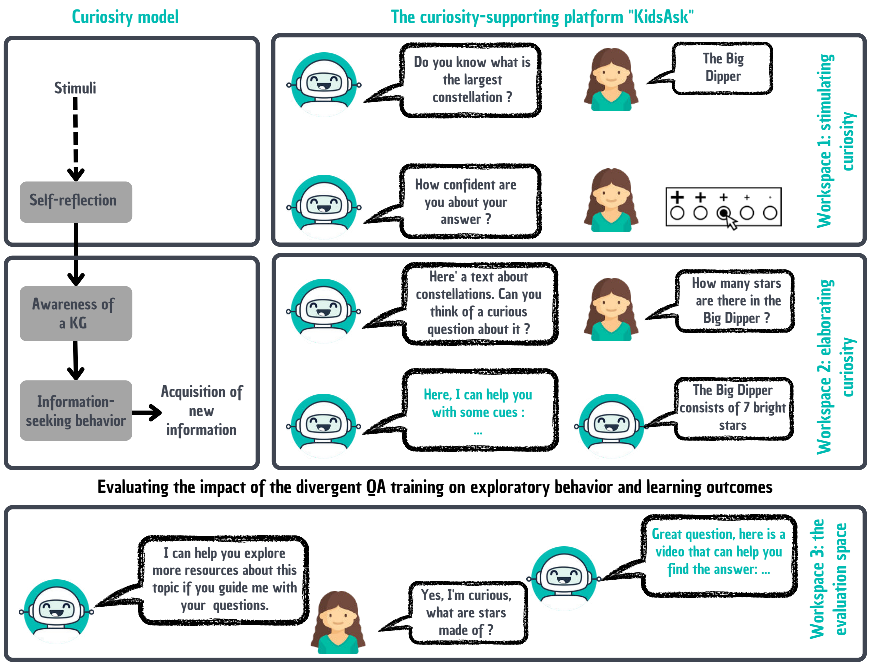


Figure 3.1: Illustration of the different work-spaces of the "KidsAsk" platform

3.2.2 *Work-spaces offered by the "KidsAsk" platform*

To meet the aims described above, "KidsAsk" offers three work-spaces: one to stimulate curiosity via encouraging self-reflection, one to elaborate curiosity via training divergent QA skills and one to evaluate the efficiency of this training in a new educational activity.

The design rationale and description of these work-spaces can be found in he following paragraphs.

WORKSPACE 1 (WS1) FOR STIMULATING CURIOSITY VIA PROMOTING SELF-REFLECTION: This space consists of a general knowledge quiz concerning six different topics (science history, animals, the environment, sports, robotics and the universe). The items of the quiz can be skipped if the child clicks on the "I don't know, I want to skip this question" button. If, however, the child submits an answer, the agent asks to report the confidence level in this answer with a 5-Likert scale, from "Super not confident" to "Super confident". We use this strategy to try to push children towards explicit self-reflection and to think deep about what they do and do not know about the question before trying to answer it. Indeed, research such as in [190] suggest that giving children the possibility to skip a question allows them to really think about what they know about the question, rather than trying to answer any how, even if randomly because they are obliged to. Adding the confidence report can also help prime curiosity furthermore as children are pushed to see information that they have some knowledge about, but not too much. This is an idea suggested by several research: curiosity rises more with a knowledge gap that has an optimal uncertainty level: not too high nor too low.

WORKSPACE 2 (WS2) FOR ELABORATING CURIOSITY VIA TRAINING DIVERGENT QA SKILLS: Children begin by choosing a topic they want to work on. They have the choice between the 6 topics they worked on in WS1. They then move to WS2 where they will practice divergent QA during a reading-comprehension task, with different texts relating to their theme of choice. For each text, they interact with the CA that helps them think of divergent questions by giving them specific cues. Depending on the experimental condition assigned to the participant (details in Section 3.3.1), these cues can either be only linguistic or a combination of linguistic and semantic cues. The semantic cues consist of pieces of information that is closely related to the text but not mentioned in it. It represents a proposition of a knowledge gap that could be interesting for children to pursue in order to deepen their understanding of the text and expand their knowledge about it. The linguistic cues are propositions of high-level questioning words that can start divergent questions. **Example: 'what difference' + 'The vaccine prevents the disease, the medication treats it' for children who have both cues and only 'what is the difference' for those who only have linguistic cues.**

The child is then asked by the agent to use the cue(s) to generate a question (for the example above, the question could be 'What is the difference between vaccine and medication?'). The agent chooses to give one or both cues depending on the condition the child is assigned to: control or experimental (see Section 3.3.1 for more details).

We choose to use this type of semantic cues as a follow-up of the study led by Alaimi et al. [8] that used these cues with CAs and showed that they resulted in children asking more divergent questions. This is linked to the idea proposed in [80, 103] suggesting that children do not ask divergent questions because they don't know that they are missing some knowledge, i.e. they don't see their own knowledge gaps. Making them explicit can thus help facilitate the QA behavior. Furthermore, we choose to offer the linguistic to help overcome the linguistic difficulties children can face when doing a question-generation task, especially with high-level questions (e.g. not being familiar with the syntactic construction of a question or how to formulate an interrogation), as shown in [80].

WORKSPACE 3 FOR EVALUATING THE IMPACT OF THE DIVERGENT QA TRAINING: In this space, children have a library of animated educational videos related to a new topic of their choice; only 3 of these are initially accessible. After every video they watch, the agent reappears to see if the video has aroused any questions in them. It also informs them that if they do ask questions, it will help them find its answer by unlocking a novel video that may contain this latter. The agent here does not give any help to formulate the questions. With every new question the child generates, the agent displays the relevant videos that are still hidden and the child picks the one they want to add to their library. Once chosen, it becomes available for playing. Children see a progress bar concerning the remaining hidden videos they can unlock if they ask questions, and can terminate the exploration phase whenever they feel they're not curious anymore. We also measure the children's learning progress in the topic with an identical quiz that children do before and after the exploration phase.

The data we collect in this space can serve three main purposes: 1) investigate if there is a "transfer" effect of the divergent QA training proposed in WS2 by analyzing children's spontaneous divergent QA skills in this new learning context where they don't have any external help to formulate their questions and are not 'obliged' to do so. 2) investigate the impact of children's spontaneous divergent QA skills on how they organise and maintain autonomous exploratory behaviors. And finally 3) investigate the impact of intensity of curiosity-driven behaviors (length of exploration, intensity of divergent QA) on the learning progress in the topic of question.

3.2.3 *Technical implementation*

The interface was programmed in JavaScript using the REACT library and was connected to a RESTful API to publish and retrieve the interaction data. As presented in Figure3.2, the behavior of the agent in terms of selection of the adequate cue(s) to offer was predefined and hand-scripted: it was connected to a database containing the different text resources and every text had a sequence of linguistic and semantic cues linked to it. Depending on the child's condition, the agent's automaton composes the dialogue utterances in order to include the appropriate support. We changed the utterances between the questions to avoid repetition in the agent's dialogue: a replica is not executed if it has been delivered during the previous question [112].

This first implementation of "KidsAsk" does not support natural language processing processing to govern the agent's behavior: the recommendation system in the exploration space was only based on an automaton that shows the resources related to the topic of choice of the child.

For both work-spaces, the questions entered by participants were only assessed later on during the data analysis phase and children had no feedback concerning their questions' quality.

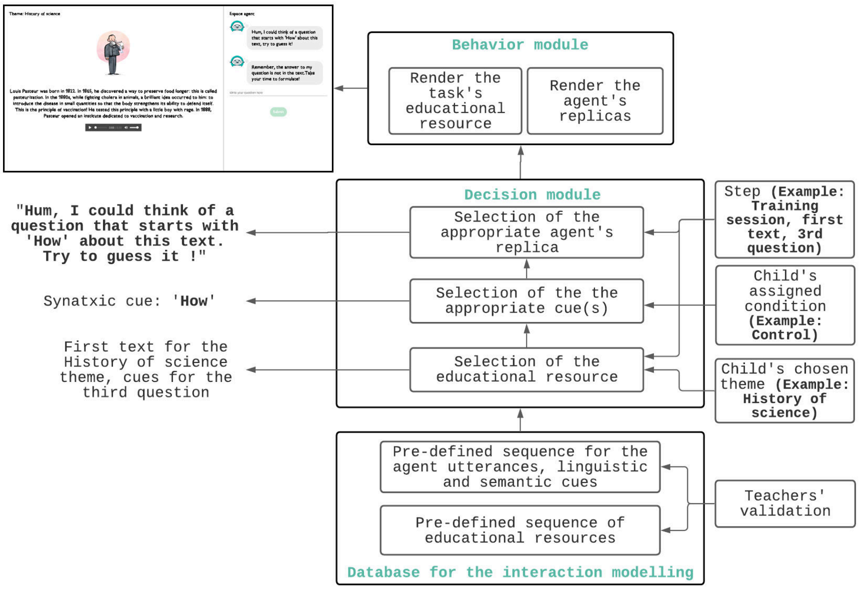


Figure 3.2: "KidsAsk" system design

3.3 EXPERIMENTAL PROCEDURE TO EVALUATE "KIDSASK"

3.3.1 *Experimental conditions*

In order to evaluate the efficiency of the curiosity-prompting approaches used by the CA, and in particular the importance of proposing knowledge gaps to help elicit divergent QA behaviors during learning, we choose to conduct an experimental investigatory study with primary-school students with two conditions like the following:

AN EXPERIMENTAL CONDITION participants interact with an agent that gives both a linguistic cue (a questioning word for a high-level question) and a semantic cue (a proposition of a knowledge gap that

Control condition	**Agent**: "I could think of a question that starts with **'What difference'** about this text, try to guess it!. **Agent**: "Remember, the answer to my question is not in the text. Take your time to formulate!"
Experimental condition	**Agent**: "I could think of a question that starts with **'What difference'** about this text. You can see the answer to it in the box below. If you would like to guess it, click on the box." **Agent:** Renders the checkbox **'The vaccine prevents the disease, the medication treats it'**. **Agent**: "Great, you can now write your question. Take your time to formulate!"

Table 1: Illustration of the difference in the CA's dialogue between the two conditions during WS2

can lead to acquiring new information). Example, for a text about Louis Pasteur's invention of vaccine (seen in Figures 3.5 or 3.6), the cues 'What causes' and 'The infection is caused by an intestine infection' are meant to lead the child to think of the question 'What causes the Cholera infection?'.

A CONTROL CONDITION participants interact with an agent that only gives the linguistic cue to help children think of divergent questions. If we take the same text example above, the cue proposed by the CA will only be 'What causes'.

In both cases, participants are free to either choose to use these propositions or think of their own question. Also in both cases, the agent does not give any feedback concerning the question entered by the child; it only acknowledges whether or its proposition was chosen by the child. See Table 1 for an example of the differences between the CA's behavior in the two conditions and Section 3.3.3 for a screenshot of these behaviors in the "KidsAsk" app.

In work-spaces 1 and 3, the agent exhibited the exact same behavior for both groups.

Measure	Control group	Experimental group	p-values
Age	9.36 ± 0.43	9.31 ± 0.43	0.66
Device use frequency	30.3 ± 6.28	27.24 ± 7.59	0.12
Curiosity trait	29.3 ± 4.3	27.12 ± 4.8	0.1
Perception of curiosity	38.8 ± 8	37.5 ± 6.58	0.53
Reading ability	280.03 ± 50.5	272.94 ± 96.6	0.75

Table 2: Profile measures for the experimental and control groups

3.3.2 *Participants*

We recruited 57 CM1 students belonging to four classes from two French primary schools, they were between 9 and 10.5 years old. We were constrained to remove the data for 6 participants who had missing or unusable data (entered incomplete or incomprehensible phrases). This left us with 51 participants that were assigned either to the control group (26 with 12 boys and 14 girls) or the experimental group (25 with 13 boys and 12 girls).

The groups were assigned with a pseudo-randomized method after collecting profile data regarding the age, the digital device use frequency, their perception of curiosity, the curiosity trait as reported by their parents and their reading score (see Section 3.3.4 for more details of the measures). Thereby, and as shown in Table 2, we had two balanced groups that were not different in terms of initial profile measures.

3.3.3 *Procedure*

"KidsAsk" was tested with children in the classroom during a four-session-long experiment. The study timeline and measures are available in Figure 3.3 and details of the different measures are described in Section 3.3.4. The study was approved by the Inria's ethics committee (certificate n°2019-23) and only started after having all participants parents' signed consents.

All sessions lasted between 45 minutes and 1h15 and were organized like the following:

SESSION 1 We started by presenting the general aim of the study and presenting the "KidsAsk" platform to participants. We then run tests and questionnaires to collect the profile measures mentioned above (more details of these measures in Section 3.3.4).

SESSION 2 Participants went to WS1 of the "KidsAsk" platform, dedicated to encourage explicit-self reflection in an attempt to trigger

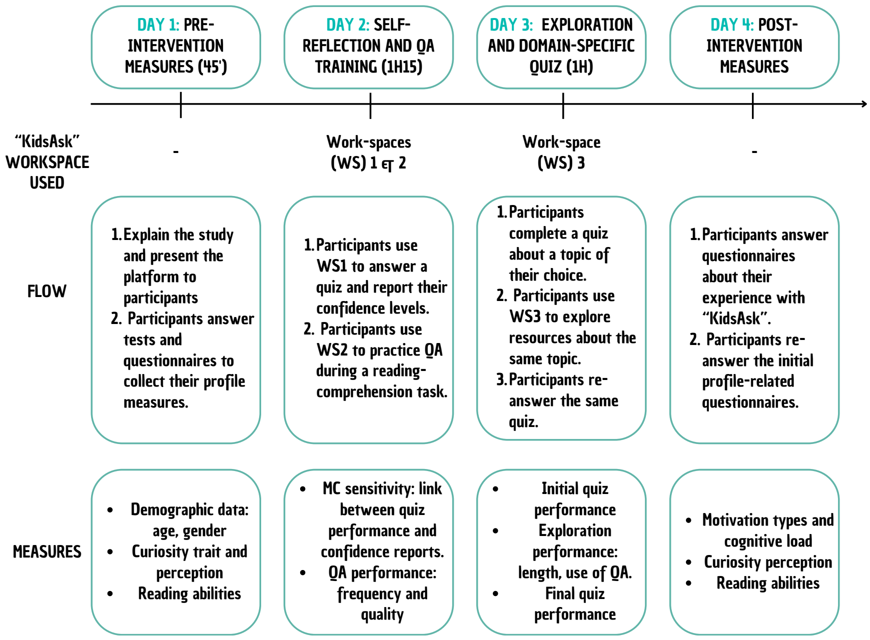


Figure 3.3: Timeline and measures to evaluate "KidsAsk"

curiosity. For this, they had a general knowledge quiz covering 6 different topics. For each question, children had the choice to either skip the item, by clicking on the 'I don't know, I want to skip this question' button, or answer it. If they do choose to answer, they are asked to report their confidence level in their answer: they had a 5-Likert scale from 'Super not confident' to 'Super confident' (see Figure 3.4). Overall, they had 18 questions.

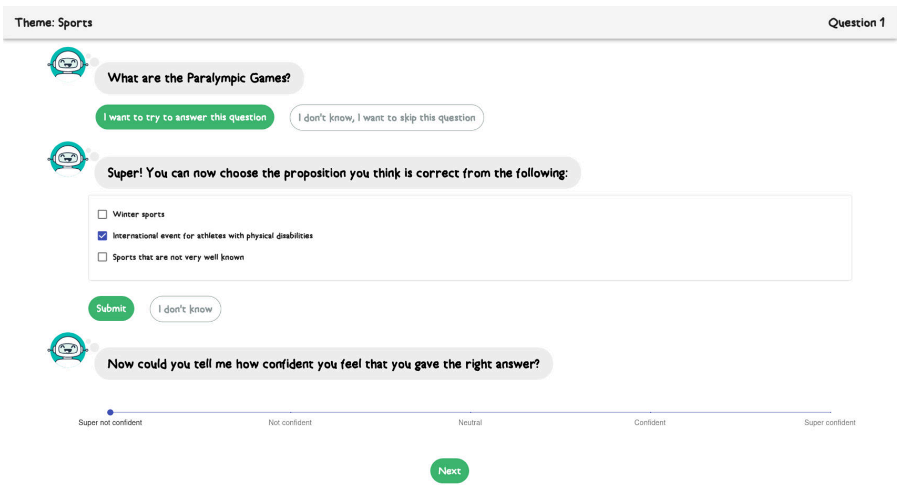


Figure 3.4: Interface for work-space 1 of "KidsAsk"

Once finished, participants choose one topic of the six they had during the quiz to work on during the divergent QA training. They then move to WS2 where they work on a reading comprehension task. During this phase, they interact with the CA that helps them with cues to formulate divergent questions about the text, and therefore, gain more information about it. To begin working on a text, participants were asked to read or listen to it (they had an audio player for each text to help with reading difficulties). Then, they click on the 'I finished reading' button once they understand it. This button enables the 'discussion' with the agent, in the agent's space on the right section of the screen. As explained in Section 3.2, the agent helps the child generate divergent-thinking questions about the text by giving either linguistic support (control group) or linguistic plus divergent-thinking semantic support (experimental group). See Figures 3.5 and 3.6 for the difference between the conditions.

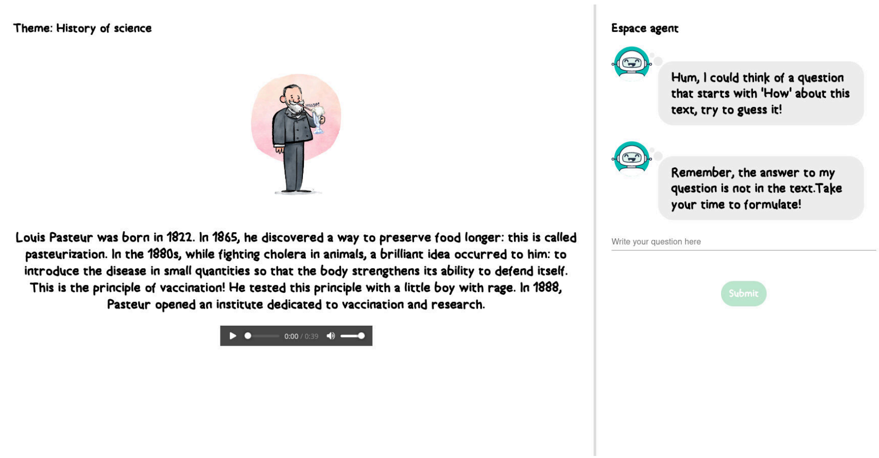


Figure 3.5: Agent behavior for the control condition in Work-space 2 of "KidsAsk"

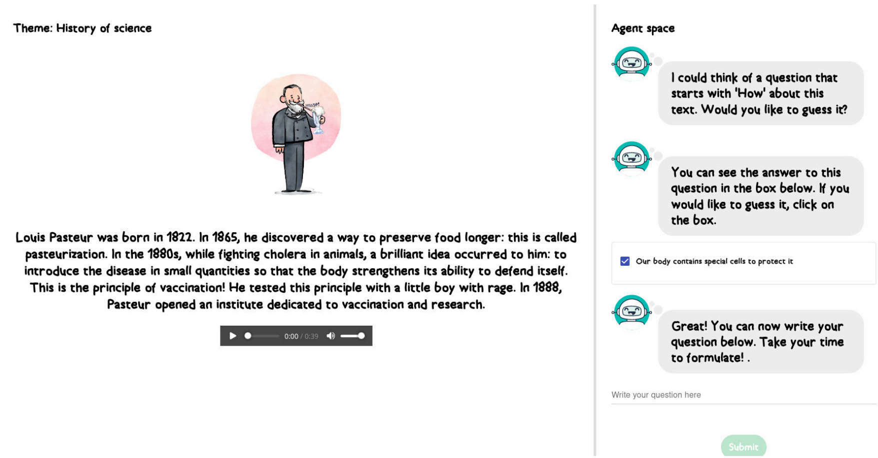


Figure 3.6: Agent behavior for the experimental condition in Work-space 2 of "KidsAsk"

The texts were selected from online resources and children magazines (Sciences et vie Junior, Quelle Histoire and Questions? Réponses!)

and were edited in order for them all to have six sentences and the same number of words. The average number of words per text was of 109. During this session, children had to process three texts and generate six questions per text with the agent, making a total of 18 questions. They were not restricted in terms of time, apart from the session length (maximum of 1h15). They had the web platform running on tablets and worked individually.

session 3 Participants start this session with a 9-item quiz concerning the a second topic of their choice. They then move to the WS3 to explore pedagogical resources about this same topic and that contain most of the answers to the questions of the quiz. During the exploration, participants are encouraged to navigate autonomously amongst several educational videos but are required to use their divergent QA skills to do so: At the beginning of exploration, participants have 3 educational videos available for watching in their 'Content table', and a progression bar suggesting that there are 6 more hidden ones. When they select a video, it appears in the central 'Content' section of the screen and they are able to play it. Once they finish playing it, they can click the 'I finished viewing' button. This button makes the agent reappear in the right section of the page. It now asks the child if they have any questions about the video they just saw, without giving any cues to help them formulate it. The agent adds that, if the child asks a question, they will have the possibility to choose a new video to unlock that may contain the answer they're looking for. If the participant decides not to ask questions, the agent has no actions and the space remains unchanged. If, on the other hand, the child submits a question, they can choose a new video to open from a list proposed by the agent. Once chosen, the video is added to the pedagogical content and becomes available for watching. The progress bar continuously indicates the remaining number of locked videos. See Figures 3.7, 3.8 and 3.9 for the agent's behavior during WS3.

There is a total of three initially-unlocked videos (children could navigate between them without generating questions), six other videos to unlock and 9 questions to generate over this session. Children were able to deliberately stop this session by clicking on the 'I finished exploring' button that appears on the top right of the screen, after opening at least three videos. The videos of this WS were all taken from the same french website 1 jour, 1 question. They were all between 1mn20 and 1mn30 of length and contained the same amount of information and the same number of sentences and drawings.

Finally, when they end this exploration phase, participants re-answered the exact same quiz they had pre-exploration, to see if they had made any learning progress thanks to their information-seeking behaviors.

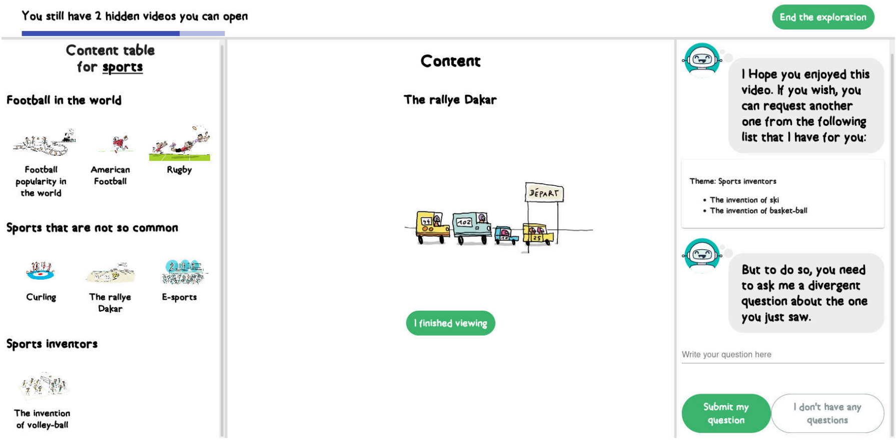


Figure 3.7: Agent behavior after viewing a video during the exploration phase in work-space 3 of "KidsAsk"

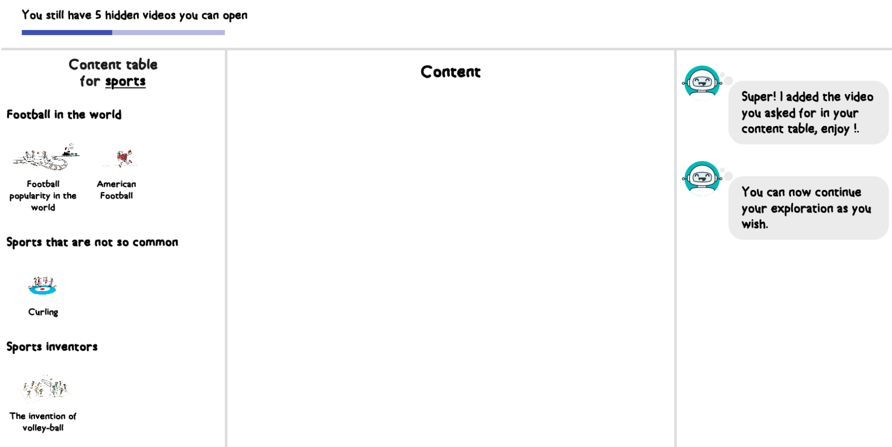


Figure 3.8: Adding a new video ("American football") to the content table in work-space 3 of "KidsAsk"

SESSION 4 Participants answered post-intervention surveys for the general motivation, types of motivation and the task load. See measures timeline in Figure 3.3 and further descriptions in Section 3.3.4.

3.3.4 *Measures*

3.3.4.1 *Profile measures*

This includes general measures: age, gender and the technology frequency use, psychological measures: curiosity trait [145] and perception of curiosity using the CIAC questionnaire [183]. And a cognitive measure: the reading abilities using the standardized test proposed in [138].

Items for the CIAC questionnaire can be found in Appendix A.2.

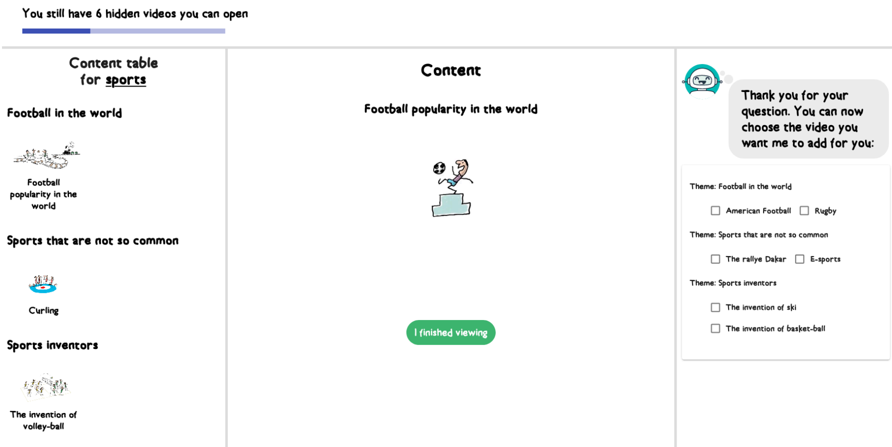


Figure 3.9: Agent behavior after submitting a question during exploration phase in work-space 3 of "KidsAsk"

3.3.4.2 *Divergent QA performance*

During the training with "KidsAsk", we collect the questions generated by participants during WS1 and WS2, i.e. when they receive support from a CA (WS1) and when they don't (WS2). We then count the number of 'acceptable' questions generated per work-space and calculate a percentage of the divergent ones.

CRITERIA FOR ACCEPTING A QUESTION A question is considered 'acceptable' if it directly related to the educational resource at hand, is not repeated and has an interrogative form. **Example:** For the text about Louis Pasteur (available in Figures 3.5 or 3.6), a linguistic cue '**What are the other**' and a semantic cue "**Fermentation**", an example of an accepted question is "What are the other ways to protect our body?" as while, it does not use the cues proposed by the CA, it still respects all other criteria. However, input such as "It is another way to preserve food" or "What are the other parts of a robot?" were not accepted as they violate one or more of our criteria.

Details about this procedure and examples of accepted and rejected input can be found in Appendix B.1.

CALCULATING THE PERCENTAGE OF DIVERGENT QUESTIONS Once a question is accepted, we evaluate whether or not it is divergent by checking if its answer is explicitly stated in the text. This is based on Gallagher's classification [69]. **Example:** The question "What are the other possible ways to preserve food?" is considered to be divergent, whereas "What are the other elements of the Pasteurisation process?" is considered convergent as the answer to it is explicitly stated in the text.

Other examples of the annotation process can be found in Appendix B.2.

SYNTACTIC QUALITY OF THE QUESTIONS: Finally, we also investigated the syntactic quality of the questions generated. Indeed, it is suggested that syntactic and semantic dimensions are tightly related: divergent questions are significantly more associated with complex syntactic constructions. With classifications such as the ones developed in [42, 69], we thus use a quotation grid that notes the question's quality depending on the syntactic construction (from 1 to 4 points given the correctness of the interrogative construction) and the use of the questioning words (from 0 to 3 points given the complexity of the questioning word used). Ultimately, questions had scores from 1 to 8 points. For each participant, we calculated an average score for all the questions generated in each work-space.

Details about this grid are available in Appendix B.4.

N.B.: All annotations described above (acceptance, divergence & syntactic quality) were performed by the two researchers who led the experiments in schools. The inter-rater reliability was of 88,1%, with an agreement percentage of 90%. All data was anonymized: coders could only see the identifiers that children were given randomly at the beginning of the intervention.

3.3.4.3 *Exploration measures*

We use the time spent in exploration in WS2 as a behavioral manifestation of participants' curiosity. Indeed, given that participants were told that they can end the session whenever they wanted, this measure can reflect on their spontaneous motivation to explore, when the only reward they get from their behavior is the information they seek. This measure is inspired from the idea introduced in [82] to have an effort-based behavioral measure of curiosity, i.e. how much are individuals willing to 'pay' for information, to satiate their curiosity.

3.3.4.4 *Learning progress measure*

Participants had to answer the same domain-knowledge quiz before and after the exploration phase (WS2); the maximum score is of 9 points. The items of this quiz were the same for all participants. They were chosen to be similar to the standard pedagogical evaluations students have in the classroom and to address the key ideas that were judged to be pedagogically important as reported by the teachers that helped us develop the content for "KidsAsk".

The quizzes contained items like the following:

- 1/3 of the questions had their explicit answers in the videos.
- 1/3 of the questions required the participant to link information from different videos to find an answer.
- 1/3 of the questions did not relate to the theme chosen for the exploration. These are our control questions to help us inves-

tigate if the learning progress made is actually related to the exploratory behavior.

The learning progress was then computed in order to evaluate the participants' progress with respect to the maximum learning gain achieved in the global sample:

$$\text{learning progress} = \frac{\text{Score}_{\text{postExploration}} - \text{Score}_{\text{preExploration}}}{\text{Score}_{\text{Max}}} \tag{1}$$

With: $\text{Score}_{\text{preExploration}}$ and $\text{Score}_{\text{postExploration}}$ being, respectively, the scores during the pre- and post- exploration quizzes. And, $\text{Score}_{\text{Max}}$ being the highest score achieved in the post-exploration quiz.

3.3.4.5 *Learner experience measures*

For the learner experience, we had two main measures: motivation and the perceived task load.

For motivation we used tow questionnaires: the general motivation scale developed in [47] and Vallerand's types of motivation in education scale developed in [223].

GENERAL MOTIVATION The general motivation scale was used to investigate the potential short-term motivation. It contains one sub-scale for evaluating participants' motivation to reuse the platform in the future, one sub-scale for evaluating their perceived competence and one sub-scale for assessing the degree of preference with respect to a favourite school activity.

The items were answered either with a 6- or 7-Likert scale, with the maximum score being 54.

TYPES OF MOTIVATION On the other hand, Vallerand's scale was used to probe intrinsic and extrinsic motivational mechanisms in our educational settings. It is composed of three sub-scales that differentiate: intrinsic motivation (possible scores from 0 to 9 points), extrinsic motivation (possible scores from 0 to 9 points) and a-motivation (possible scores from 0 to 3 points).

All questionnaire items are yes or no questions.

TASK LOAD Finally, we also measured participants' subjective workload using the NASA-TLX workload multi-dimensional scale developed in [92]. The measure contains six dimensions: mental demand, physical demand, temporal demand, performance, effort and frustration. The maximum score is 60.

Items for these three questionnaires can be found in Appendices A.4, A.5 and A.6.

3.4 RESULTS

In order to address our research questions, we conduct our data analysis specifically to investigate:

1. The efficiency of our approach in helping children ask more divergent questions. More particularly: does adding semantic cues that represent potential knowledge gaps helps children ask more divergent questions during WS2?

2. The transfer effect of the divergent QA training in a new learning context and its impact on spontaneous exploratory behaviors: when they change to a new learning context (i.e. in WS3), can children still exhibit divergent QA skills? And how does this affect their exploratory behaviors?

3. (how) Do curiosity-driven behaviors (exploration, divergent QA) impact children's domain-knowledge learning progress?

3.4.1 *Efficiency of the QA training and transfer effects in WS2*

We start by studying the impact of our training on children's divergent QA behaviors both during the training where they had the support of an agent to do so (i.e. in WS2) and during a new learning task where they had no incentives or external help to formulate their questions (i.e. WS3).

3.4.1.1 *Efficiency of the divergent QA training in WS2*

In analyzing the efficiency of the training given by the two agents during WS2 (i.e. giving linguistic cues vs. giving linguistic cues + propositions of knowledge gaps), we compare children's performance with respect to two dimensions: the percentage of divergent questions they were able to generate and their syntactic quality. Both dimensions were quoted by hand using the grids and procedures explained in Section 3.3.4.

As expected, children from the experimental group (i.e. who had help also concerning the identification of knowledge gaps) ended up formulating more divergent questions during WS2 that the ones from the control group where the agent only gave linguistic help ($t=-4.19$, $p=0.0001$).

We also investigated the syntactic quality of these questions. This is indeed an important measure for our study since syntactic quality is shown to be tightly correlated to the divergence level of questions [80]. We use the grid described in Section 3.3.4 and we calculate an average score. Similar to the previous analysis, we find a significantly better performance for participants from the experimental condition ($t=-4.36$, $p<0.0001$). See Figure 3.10 for more details.

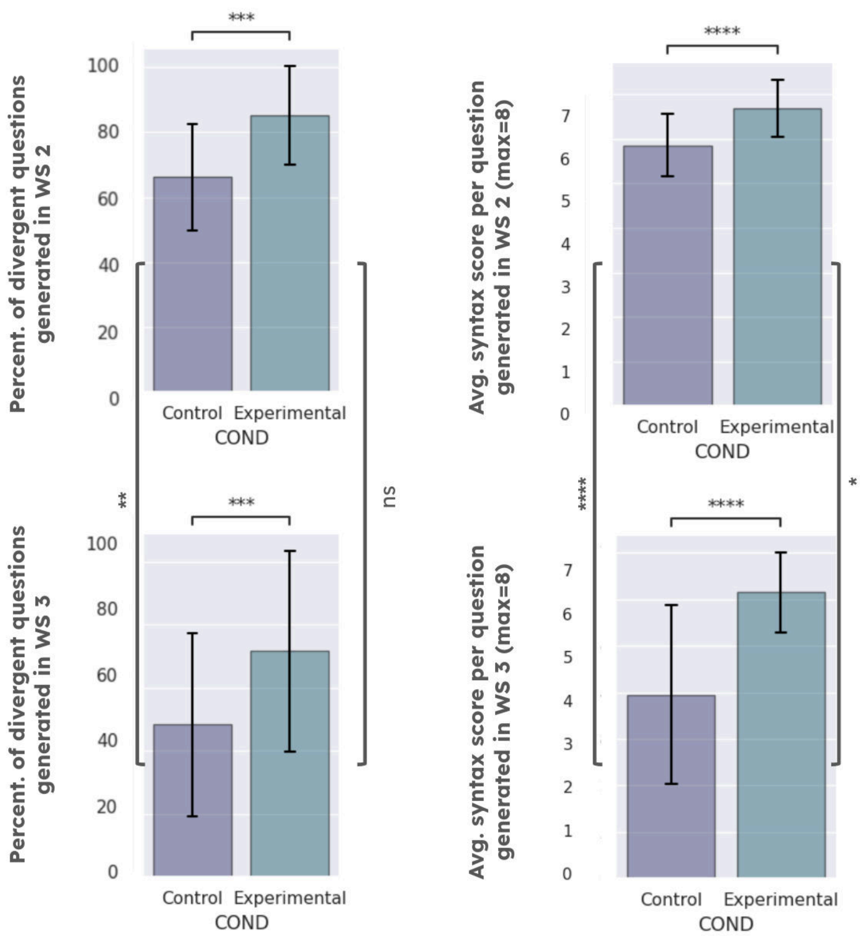


Figure 3.10: Participants from the experimental group generated significantly more divergent questions and with better syntactic quality in both work-spaces of "KidsASk". Their performance dropped with a smaller effect when they moved to WS3 (i.e. with no external support or obligation to generate questions), revealing a stronger efficiency for the experimental agent in transferring the divergent QA skill.

3.4.1.2 *Transfer of the divergent QA skill in WS3*

After verifying the efficiency of our incentive agent in helping students ask more and better divergent questions during WS2, we also investigated if we can find a transfer effect of the training in WS3, where participants are in a new learning context and have no external support or 'obligation' to generate divergent questions.

Our first results revealed that children from the experimental group were significantly more successful in generating more divergent questions (t=-2.67, p=0.001) and with better syntactic quality (t=-5.13, p< 0.0001). Furthermore, we also run a mixed ANOVA analysis to investigate the effect of the experimental condition and the work-space nature on participants' QA performance in its two dimensions. As shown Figure 3.10, the results revealed a significant effect both on the percentage of the divergent questions generated ($F(1,49)=10.87$, $p=0.001$) and their quality ($F(1,49)=32.12$, $p<0.0001$).

The post-hoc tests investigating the simple effect of the work-space for each condition on the percentage of divergent questions generated revealed a significant interaction only for the control group ($p=0.009$); the performance dropped drastically. This was not the case for the experimental group ($p=0.06$) where the performance tended to stay stable. However, concerning the syntactic quality of the questions, the performance dropped significantly for both groups when participants moved to the exploration space (i.e. where they had no support to generate questions). The drop was more drastic for the control group ($p<0.0001$) than for the experimental group ($p=0.01$) (See Figure 3.10).

Taken together, these results suggest that training children to ask divergent questions is more efficient when helping them discover potential knowledge gaps. With such a training, participants were also more able to use the divergent QA skills they learnt spontaneously an in different learning contexts, where no external help is given to help them.

3.4.2 *Exploratory behavior and link with the divergent QA performance*

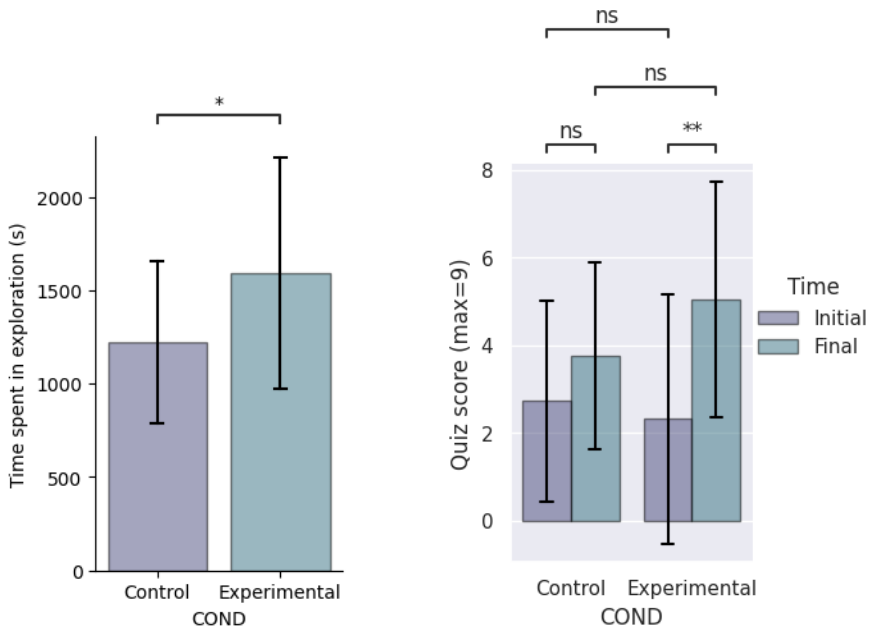
In a second step, we wanted to investigate the effect of the divergent QA training given in WS2 on participants' autonomous exploratory behaviors. To do this, we analyze the time participants spent exploring the educational resources spontaneously. This can be seen as a spontaneous curiosity-driven exploratory behavior since children had the possibility to end the session whenever they decided they are not interested in exploring any more new content. Results showed that participants from the experimental group ended up spending significantly more time exploring the educational resources compared to the control group ($t=-2.42$, $p = 0.01$ and an effect size: Cohen's d of 0.68). See Sub-Figure (a) in Figure 3.11.

We also run an ANCOVA test to analyze the impact of the divergent QA training given in WS2 on this measures. We find indeed a significant interaction: $F=4.1$ and $p=0.04$.

Together, these results suggest that the more participants were able to benefit from the divergent QA training they had during WS2, the more they were able to lead and maintain autonomous explorations.

3.4.3 *Domain-knowledge learning progress*

In analyzing the two groups' learning progress, we compare participants' performances in the domain-knowledge quizzes before and after the exploration. It is to be noted that children's scores were ini-



(a) Participants from the experimental group spent significantly more time exploring educational resources in WS3.

(b) Participants from the experimental group had a significantly higher learning progress

Figure 3.11: Impact of "KidsAsk" on exploratory behavior and learning progress

tially similar for the two conditions, meaning that they both had the same space to achieve progress.

To do this, we run a two-way ANOVA analysis to see if the difference between the initial and final performances was significantly different between the two conditions. As shown in Sub-Figure (b) in Figure3.11, the results showed a significant effect of the time (F(1,49)=32.9, p=<.0001) as well as the interaction between the two (F(1,49)=4.31, p=0.004). Hypotheses concerning the absence of outliers and normality of the learning progress variable were verified before running the test.

Finally, in analyzing the predictors of the learning progress across the two groups, we find a significant effect of the divergent QA performance during the exploration (F=3.67, p=0.05). These results support previous research results where learning is enhanced with curiosity-driven information-search behaviors which at their turn in our study, where enhanced with the divergent QA training proposed in WS2. See Figure 3.12 for a summary of the relationships found between our different measures.

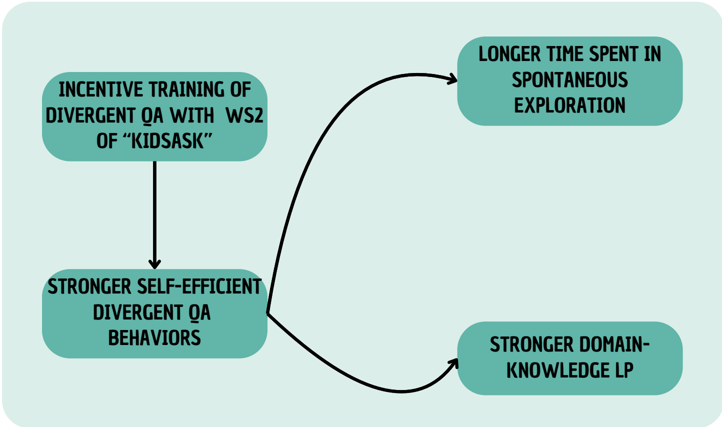


Figure 3.12: Summary of the results seen with the "KidsAsk" training

3.4.4 *Learning experience measures*

3.4.4.1 *Motivation*

The scores of general motivation did not differ between the two conditions (p=0.54) but remained high throughout the two sessions (m=39.3 for the experimental and m=37.74 for the control; max score =54). The difference between the groups in terms of intrinsic motivation was not significant either. However, and as shown in Figure3.13, participants were significantly more intrinsically than extrinsically motivated for both groups and during both training sessions with "KidsAsk".

3.4.4.2 *Task load*

Participants from the experimental group tended to report lower cognitive load feelings after using "KidsAsk" (t=1.86, p=0.06). These results may suggest that the task of generating divergent questions, either with or without external support, was perceived to be slightly easier for participants who had the incentive agent i.e. an agent that facilitates seeing knowledge gaps knowledge gaps).

3.5 DISCUSSION

In this first study, we investigate the effectiveness of implementing a conversational agent designed to help students ask more divergent questions by proposing knowledge gaps and offering linguistic support to transform them into questions. We also investigate the im-

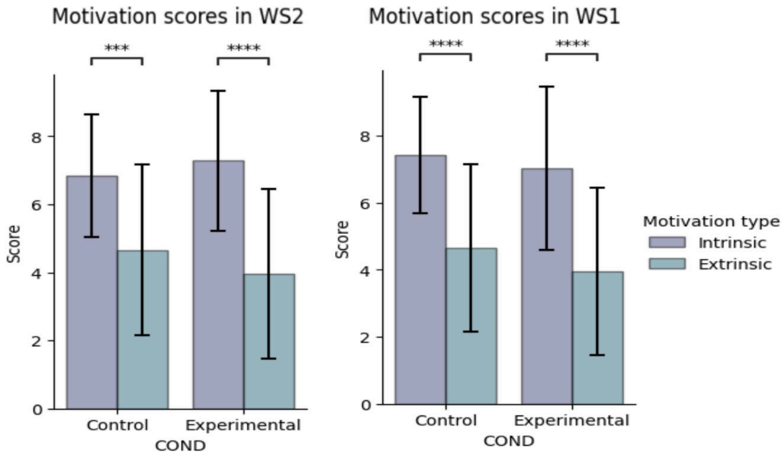


Figure 3.13: Participants from both conditions were more intrinsically than extrinsically motivated when using "KidsAsk"

pact of such training on fostering children's ability to spontaneously ask divergent questions while learning, conduct autonomous explorations and use these strategies to make learning progress.

The key idea behind our design is coming from the "knowledge gap" theory, i.e. individuals will be more curious and ask questions when they realize that they are missing some piece(s) of information. In this context, we designed an 'incentive' agent to facilitate this realization by encouraging self-reflection in a first step and proposing potential knowledge gaps to investigate in a second one. To validate this design, we compare this agent with a 'neutral' one that only encourages self-reflection in the beginning of the interaction but does not propose knowledge gaps to facilitate question-asking.

As expected, our findings suggest that participants who interacted with the incentive agent were more successful in generating divergent questions during the training session. These participants also ended up asking more and better divergent questions during the exploration phase, where they had a different learning context (different domain topic and different type of task) and no external support to formulate their questions; they were also not obliged to ask questions in this phase. We thus can suggest that this transfer effect of the divergent QA training may be due to the incentive agent's success in: 1) helping children do more self-reflection and be at ease with identifying knowledge gaps they want to pursue, and 2) helping them have the linguistic skills allowing them to pursue these gaps using well-formulated divergent questions. These findings are inline with previous research such as in [80] suggesting that children need both linguistic support and self-reflection/ awareness to boost their QA behaviors.

In addition, our results suggest that the group that interacted with the incentive agent ended up spending more time in exploration. As discussed above, we consider this measure as a manifestation of curiosity to learn since participants had no obligation to finish the task and the only rewards they received from their exploration was the information they seek itself. Our results also suggested that this measure depended on the divergent QA performance during the exploration, i.e. the more participants asked divergent questions, the more time they spent exploring. This finding can be compared to the results found with "think aloud" approaches for instance. Indeed, these approaches consist of verbalizing one's thinking process explicitly while trying to solve a problem and are shown to lead to more carefully-planned problem-solving strategies [16, 96]. This may suggest that training the faculty to stop, think of and formulate divergent questions while learning might be a valuable strategy to be fostered in children in order to prevent them from the 'illusion of knowledge trap' and from prematurely interrupting their information-seeking cycles. Indeed, several studies suggest that children tend to end their learning cycles prematurely because they overestimate their mastery of the skills [129, 168].

As for the domain-specific learning progress, we found that the experimental group was also more successful. We also find a correlation between this learning progress and the divergent QA performance during the exploration phase, i.e. the more divergent questions were asked, the more learning progress was made. This finding reinforces work such as in [175] suggesting a causal and bi-directional link between curiosity and learning. It is also in line with studies such as in [117], where authors show that states of curiosity can enhance learning and memory retention. With the results we have so far, we can therefore suggest that getting children familiar with the process of generating appropriate divergent questions to gain new knowledge helped them do more explorations and, thus, make more learning progress.

As a final important observation, our results failed to show higher intrinsic motivation scores for the experimental condition. This is counter our hypothesis (inspired from the LP theory [176]) that children with the more learning progress will report more intrinsic motivation. Several explanations can be advanced. First, and as mentioned in [79], the learning by asking questions strategy, especially using digital tools is still rarely used in classroom settings, even today. It is therefore possible that the activity's novel and playful characters alone have made it very motivating and attractive for children, with both agents. Another possible explanation is children's incapacity to accurately self-assess their learning progress, due to their metacognitive immaturity [155]. Indeed, learning judgment failures would reduce the reinforcing power of learning progress on intrinsic motivation.

Finally, we can suggest that with rather encouraging results, our incentive curiosity-prompting approach succeeded in promoting divergent QA; a skill that helped participants shape their autonomous exploratory behaviors and make useful learning progress.

3.6 LIMITATIONS AND FUTURE DIRECTIONS

A main limitation to this first study is the low intelligence level of the CA. Indeed, as discussed in the design section above, the current implementation does not support any kind of natural language processing methods during the interaction with the child. For this reason, and seeing the encouraging results that suggest the validity of our approach, we will pursue our work with a perspective to endow our agent with higher computational powers using large language models. With this new implementation, we aim to propose an agent that can provide personalized feedback on the questions that children generate (convergent vs. divergent, syntax errors ..). We also aim to implement an agent that can extract the relevant semantic cues that can elicit children's curiosity, automatically from the educational resources. Indeed, this process has been very challenging until now as it was done by hand and was rather time-consuming.

Another factor to consider is the influence that the study context may have had on our results. Indeed, our study was conducted in a supervised classroom where members of the research team were present to address the participants' inquiries and help with any difficulty they faced. Also, the less-formal presentation of the resources (audios, videos) could have formed favourable conditions for our incentive agent to work. This could give us general directions to replicate our work in different classroom settings, without the presence of a research team and with more formal educational activities.

Finally, it is also to be noted that one of our aims is to raise children's awareness about the importance of the curiosity-driven behaviors on their learning. But with this experiment, and even though we were able to see an improvement in these behaviors and in learning progress, we do not know if the intervention actually helped participants see the link between the two. A future direction for this work will be to add a space in the platform where users can find pedagogical tutorials and activities that explain how to use and control their curiosity-driven behaviors to boost their learning.

3.7 CONCLUSION

In this first study, we contribute to the promotion of curiosity approaches that foster inquiry-based learning, while exploring the social advantages of using artificial conversational agents. We show that curiosity-driven behaviors such as curious QA and exploratory information-searching can be practiced with the agent, and that these behaviors are related to the learning progress children can achieve.

While our design still presents various limits, our positive findings motivate the implementation of such approaches both in classroom settings and e-learning environments. They also suggest that monitoring divergent QA skills could be a valuable indicator for teachers to prognosis their students' learning progress.

4 STUDY 2. TOWARDS THE AUTOMATION OF CURIOSITY-DRIVEN BEHAVIORS TRAININGS: ROLE OF GENERATIVE AI

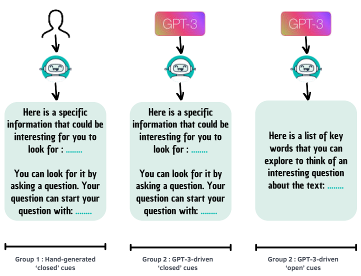
Collaborators: Xingdi Yuan & Tong Wang, Microsoft Research Montreal. Yen-Hsiang Wang, National Chung Hsing University, Taiwan.

Aims: This chapter aims to study the feasibility of using GPT-3 to generate the pedagogical content for a divergent QA training similar to "KidsAsk".

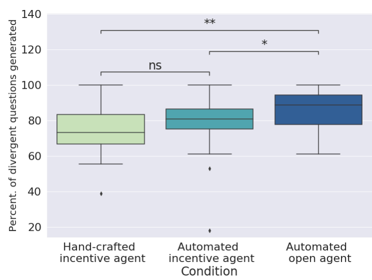
Abstract: The chapter introduces two design methods to generate the pedagogical content of "KidsAsk" using GPT-3, i.e. two families of cues proposed by the agent to facilitate divergent QA. It then presents a study with 75 students aged 9 to 10 to test the efficiency of these cues. They had interactions with an agent that either provided: 1) hand-generated "closed" cues leading to a predefined question, 2) GPT-3-generated content with this same structure, or 3) GPT-3-generated "open" cues leading to multiple questions. Finally, the chapter presents the results for the quality of GPT-3's output and impact on children's divergent QA. Using comparisons with hand-generated cues, results suggest that GPT-3, especially with "open" cues, can be efficient for training divergent QA.

Contents

The work presented in this chapter is published in IJAIED, 2023: ***Rania Abdelghani et al. "Gpt-3-driven pedagogical agents to train children's curious question-asking skills." In:*** **International Journal of Artificial Intelligence in Education** ***(2023), pp. 1–36***



(a) "KidsAsk" implementations we tested with and without GPT-3



(b) Impact on divergent QA

4.1 INTRODUCTION

During the last chapter, we showed that our methods used in "KidsAsk" to train children's divergent QA abilities (i.e. by designing CAs that propose relevant questioning words and possible knowledge gaps to pursue) were rather effective in fostering this behavior even in different learning contexts, in fostering autonomous exploration and in boosting learning [3, 8].

However, and as pointed out in the limitation of this study, the implementation of "KidsAsk" is still very limited, mainly due to the difficulties associated with generating the educational content governing the CA's behavior. Indeed, during our first implementation of "KidsAsk", we had a manual process to generate all the cues proposed by the agent for the different tasks, which revealed to be a very difficult and time-consuming. This is, in fact, part of a broader challenge that faces human teachers as well as EdTech designers in general: having to continuously hand-generate large data-sets of pedagogical content can be a very redundant, time-consuming and a costly task.

In this context, we propose to leverage advances in the natural language processing (NLP) field and investigate the efficiency of using a large language model (LLM) to partially automate the production of content for this specific pedagogical task, i.e. divergent QA training. More particularly, we use the GPT-3 language model to generate specific linguistic and semantic cues that can help children think of and formulate divergent questions about a task at hand, similar to the cues generated manually in the "KidsAsk" study. For this, we use a method called "prompting" that consists of describing a task in natural language to the LLM in order to generate the wanted output [34]; in our case, this output is the cues are the semantic and linguistic propositions used to help children think of divergent questions.

We investigate two different approaches to generate these cues: 1) one to generate cues with identical structure to the those generated manually in "KidsAsk". 2) one to generate a new type of semantic cues, that can help children have more choice over the questions they can formulate, as opposed to the approach in "KidsAsk" where the cues lead to think of one specific, predefined question. We use human expert annotations and comparisons with hand-generated content to assess the quality of these GPT-3-generated cues. We also conduct a field study (with 75 children aged 9-10) and use human annotations to evaluate children's divergent QA performance when having this training with hand-generated content vs. the two types of GPT-3-generated one.

Concretely, we proceed like the following:

1. In a first step, we use GPT-3 to generate divergent QA-eliciting cues given specific educational resources (text-based) and compare them to what human experts can produce when given the same task. We test generating two types of cues : 1) similar to what was implemented in "KidsAsk": incentive 'closed' cues that are composed of a short sentence to guide children towards thinking about one specific relevant information they might be missing about the task at hand (a proposition of a knowledge gap)+ a questioning word that they can use in order to formulate a divergent question concerning this identified missing information. 2) 'open' cues, in the form of important keywords about the text that can lead children to recognize their own missing information and ask the questions they want with more freedom (as opposed to being constraint to find the question concerning one specific piece information) + a list of questioning words they can use.

   We then use human experts annotations and comparisons with hand-generated content in order to evaluate the output of GPT-3 in terms of semantic relevance to the text, divergence level and syntactic quality.

2. In a second step, we conduct a field study with 75 children aged 9-10 to evaluate their divergent QA performance when having these different kinds of cues. To do this, we also use human experts annotations to compare the questions they formulated, in terms of divergence level and syntactic quality, when interacting with a conversational agent that proposes incentive 'closed' cues generated by hand (incentive hand-crafted CA) vs. generated by GPT-3 (incentive GPT-3 CA) vs. when having 'open' cues generated by GPT-3 (open GPT-3 CA).

3. Finally, we evaluate the effect of the training (in its three different forms) on: children's spontaneous divergent QA behaviors in a new learning context, their perception of their own competency in this skill and their attitude towards curiosity and asking questions in general.

This investigation seems crucial to us prior starting to use LLMs in educational technologies and/or by teachers to generate pedagogical content, as they can still show poor performances despite their impressive powers (e.g. reasoning and cognitive biases, hallucination problems, etc [122, 124]).

4.2 IMPLEMENTING "KIDSASK" WITH GPT-3

Given the aims of the current study, we decide to only use two of "KidsAsk" work-spaces: one to stimulate curiosity via encouraging self-reflection (WS1) and one to train divergent QA skills using the cues proposed by the CA (WS 2). During WS2, and depending on the condition assigned to participants, the cues proposed by the CA can have an 'open' or 'closed' structure and are either generated by hand or by GPT-3 (more details about these different implementations and their technical properties are described in the upcoming sections).

The work-spaces we use are the two first ones illustrated during the previous chapter to present "kidsAsk" (Figure 3.1). These work-spaces are:

WORK-SPACE 1 (WS1) TO ELICIT CURIOSITY VIA ENCOURAGING SELF-REFLECTION As for "KidsAsk", this WS contains a general knowledge quiz about different topics and pushes participants to explicit self-reflection after every quiz item. It does so by proposing to skip the question if they are sure they don't have the answer and to report a confidence level in their answer if they end up answering.

WORK-SPACE 2 (WS2) TO ELABORATE CURIOSITY VIA TRAINING DIVERGENT QA SKILLS Similar to "KidsAsk", children complete a reading-comprehension task and interact with the CA that helps them think of divergent questions to better understand the text, by proposing specific linguistic and semantic cues. The linguistic cues are propositions for high-level questioning words that children can use to formulate divergent and complex questions. Semantic cues can be: 1) a proposition of a knowledge gap concerning the text in the form of a short sentence, that was generated by hand (hand-crafted incentive CA) , 2) a proposition of a knowledge gap concerning the text in the form of a short sentence, that was generated by GPT-3 (GPT-3-driven incentive CA), or 3) two important key words that are related to the text, generated by GPT-3 (GPT-3-driven open CA). See Section 4.2.1 for more details about the choice of the structure for these cues.

Children have one of these three types of semantic cues depending on the condition they are assigned to: hand-crafted incentive agent, GPT-3-driven incentive agent or GPT-3-driven open agent. For these three conditions, the agent does not give any feedback concerning the question formulated by the child.

4.2.1 *Experimental conditions*

Our aim is to evaluate GPT-3's efficiency in generating cues that can facilitate divergent QA skills in children. To do this, we aim to generate cues of the exact structure both by human experts and GPT-3 and then compare their quality and their effect on children's divergent QA abilities. We also aim to use this study to investigate a new structure of these cues where we give children more space to think of their own questions. To meet these aims, we set up three experimental conditions:

hand-crafted incentive agent (group 1) A control group where the CA is controlled by human-generated incentive cues that lead to predefined specific divergent-thinking questions, i.e. the proposition of one specific knowledge gap (in the form of a short sentence) that could only lead to asking one specific question. We choose to maintain this strategy —identical to what was used previously in "KidsAsk" [3], given the positive effects it had on training children's divergent QA skills and exploratory behaviors.

automated incentive agent (group 2) A first test group where the CA is controlled by GPT-3. It is designed to replicate the exact same type of cues proposed by hand for the control group : participants have GPT-3-generated incentive cues that lead to predefined specific divergent-thinking questions. Since we saw positive results when produced by hand, we want to see if we can have similar effects when using GPT-3 to generate them. We therefore use specific prompting approaches of the LLM to generate this specific type of divergent QA-inciting cues (see Section 4.2.2 for details about the prompting strategy).

automated open agent (group 3) A second test group where the CA is controlled by GPT-3-generated "open" cues (i.e. important key words to guide attention towards important concepts of the text) that can lead to several possible divergent questions about the text at hand. The motivation of this type of cues is to leave children with more choice over the questions they can formulate during the training. Indeed, having heavily teacher-directed tasks that impose specific tasks to all children (like the case for our first type of cues) may alter with their sense of control and autonomy which are central factors for motivation and curiosity according to the Self-Determination Theory [86, 197]. Furthermore, proposing the same support for all children is likely to result in giving trainings that do not match with their different competency levels and zones of proximal development (ZPD). Therefore, giving cues that lead to the same predefined specific questions identified by the expert can lead children to ask questions that are too complicated for them or, on the contrary, questions that they already know the answer to. This can mean that the training is likely not to support the users' curiosity since this latter is tightly related to each individual's ZPD and sense of competency [159]. Finally, we also think that it is important to give different choices in order to avoid the cases where the human experts (or the LLMs') choice of the semantic cues is heavily influenced by their preconceptions and beliefs about children's knowledge gaps [115, 222].

Therefore, we propose propose to use GPT-3 to extract relevant keywords about a given text. These keywords can be used by the CA to raise children's awareness about the important concepts of the texts that need further investigations, and guide them towards thinking of different divergent questions. Investigating children's divergent QA performance when having these GPT-3-generated open vs. specific cues (i.e. Group 2 vs. Group 3) could therefore help us investigate different contexts where children's curiosity can be induced more efficiently.

See Figure 4.1 for details about the difference between the agents' dialogues and implementations for the different conditions.

The agent's behavior in WS1 (i.e. during the self-reflection phase) is identical for the three groups.

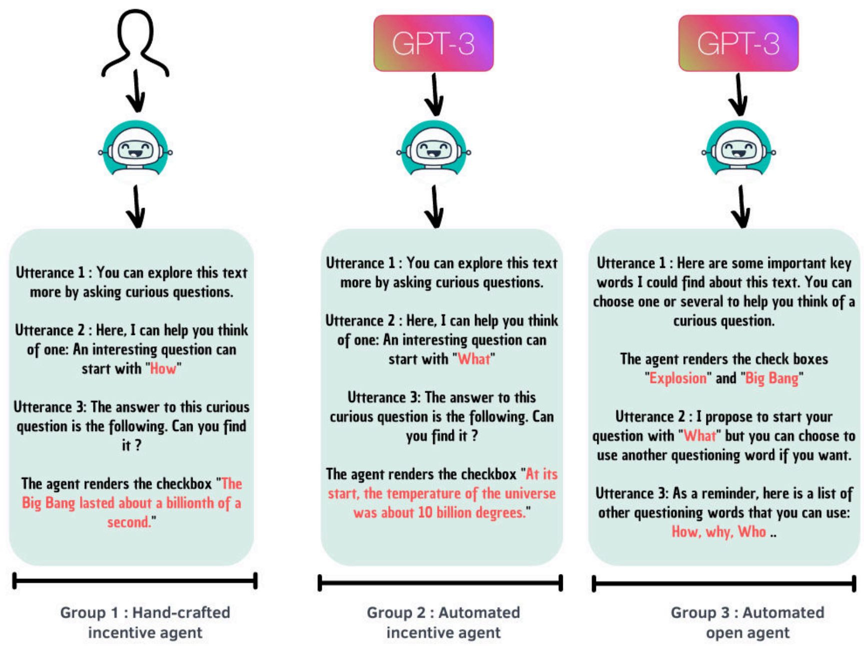


Figure 4.1: Difference between the agents' behaviors and implementations for the three experimental conditions

4.2.2 *Technical implementation*

4.2.2.1 *Choice of the technology*

We choose to work with the pre-trained LLM GPT-3 and to use prompt-based methods for two main reasons. First, at the time of conducting this study, GPT-3 was one of the most state-of-the-art and robust LLMs, with a training upon text corpus containing the widest variety of topics [34]. Moreover, it has demonstrated robust performances in various downstream tasks, and more importantly to this study, in knowledge-based question-asking and question-answering tasks and coding of curiosity-driven questions [229, 231].

Second, and with the aim to implement an easy-to-use system that can be accessible for the broad teaching community, we choose to work with GPT-3 in a prompt-based methods setting [34]. Indeed, and as discussed above, prompt-based learning consists simply of giving pieces of natural text to the model as an instruction in order to adjust it to the specificities of the task in question and generate the output needed (in our case, this output is the set of cues as described in Section 4.2.1). So with this simple implementation, we avoid complex methods such as fine-tuning that require a machine learning expertise and collecting a large set of specific data, etc. Instead, we provide a system that can be easily used by practitioners (e.g. teachers) and/or adapted to different activities of their choice (e.g. change the language, target a different type of questions ..), without the need for a specific expertise.

4.2.2.2 *Generating the cues with GPT-3*

For our experiments, we prompt GPT-3 in a zero-shot setting [34], meaning that the model generates output based only on a context and natural language description of the task, i.e. a prompt, without giving it any examples of the desired output. Concretely, we provide a dataset containing the text-based educational resources: short scientific articles with a mean of 109 words per article, that will serve the model as context. Our task consists of building the proper prompts that can lead GPT-3 to produce a questioning word combined with either a sentence that represents a proposition for a knowledge gap (for the group 2) or a pair of relevant keywords (for the group 3) that can be used as cues to formulate a divergent question relevant to the context. See Figure 4.2 for details about the prompts approaches we retained.

Before deciding on the prompting approaches we ended up using, we run several trials where the following configurations were tested :

- Story then instruction vs. instruction then Story: we tried a set of different ways to formulate the prompts, including the ordering of story and the instruction. To do this, we highlighted

[Story]:
The bees perform a dance when they want to indicate to their colleagues the position of a nectar source. They perform a figure eight choreography on the vertical combs of the hive.
If the bee moves up and down the figure of eight, it means that it should face the sun; if it moves in the opposite direction, it means that it should face away from the sun.
Bees also communicate by means of "vibrations": the "grooming" vibration, for example, which a bee emits when it needs a colleague to clean it, or the "beggar" vibration. The bees also do a semi-circular dance.

[Instruction]:
For the above [Story] , generate 5 [Key ideas].

[Key Ideas]:
1. Bees use dance to communicate the position of nectar sources.
2. Bees communicate by means of vibrations.
3. The figure eight choreography is used to communicate the position of the nectar source.
4. The semi-circular dance is used to communicate the direction in which the bee should face.
5. The "grooming" vibration is used to communicate the need for cleaning.

[Instruction]:
For each of the [Key Ideas] above, generate 2 related words on the same topic.

[Related Words]: **// Used as the semantic cues for the automated open agent**
1. Bees: honey, pollination
2. Vibrations: communication, signals
3. Figure eight: choreography, dance
4. Semi-circular: dance, communication
5. Grooming: cleaning, hygiene

[Instruction]: Use the above [Related Words] to ask 5 divergent questions.

[Divergent Questions]: **// The questioning words are used as linguistic cues for the two automated agents**
1. What other ways do bees communicate?
2. What other functions do vibrations serve in bee communication?
3. What are other examples of figure eight choreography in nature?
4. What are other purposes of the semi-circular dance?
5. What are the consequences of not grooming?

[Instruction]:

Answer the above [Divergent Questions] using sentences with less than 10 words.
[Answers]: **/ /Used as the semantic cues for the automated incentive agent**
1. Bees also communicate by smells and tastes.
2. Vibrations also serve to warn of predators.
3. Other examples of figure eight choreography include birds flying in formation and fish swimming in schools.
4. The semi-circular dance can also be used to communicate the direction in which the bee should fly.
5. Not grooming can lead to infection and disease.

Figure 4.2: Prompting approaches used to control the agents' behaviours

which part of the prompt is the story and which is the instructions by adding the specific indicators (i.e. "Story:" and "Instruction:"). We also tried different formats for the prompts by choosing to add, or not, a new line after every new indicator. We did not observe significant difference from these variants, probably because LLMs at the scale of GPT-3 are sufficiently robust to bypass these dimensions. We thus decide to stick with a format (as shown in Figure 4.2) that is considerably straightforward, readable and re-usable for practitioners.

- Zero-shot vs. One-shot vs. Few-shots vs Fine-tuning: In all the experiments described in this work, we prompt GPT-3 in a zero-shot manner. The reason is two-fold. 1) several recent work in the NLP community reports that pre-trained LLMs such as GPT-3 can be sensitive to the specific data points provided as few-shot examples [132, 148, 163, 234]. For instance, LLMs are observed to suffer from "recency bias", i.e., they overly rely on the examples that are located closer to the end of the prompt and thus tend to bias the output towards copying from the most recent examples. Despite many works have been proposed to alleviate such issues, they are still recurrent with GPT-3. And 2) we want to emphasize that in most real-world scenarios such as pedagogical applications, practitioners are less likely to have machine learning expertise nor sufficient computational resources to fine-tune an LLM, or to optimize the LLM outputs by providing the model with different combinations of examples. We envision that the proposed system can have greater accessibility if it does not require too much expertise from the users.

In all experiments, we use the text-davinci-002 variant of GPT-3 and a temperature of 0.7 during the prompting process.

4.2.2.3 *Connecting "KidsAsk" to GPT-3 output*

As we have no control over GPT-3's outputs, we choose not to put children in direct contact with the model to avoid any potential unfair interactions. For this reason, we opt for an offline implementation like the following:

- We use GPT-3 to generate the agents' cues like described in the section above.
- We repeat the procedure several times for the same resources and we take some of the outputs randomly as cues to appear in the platform.
- After verification that these cues do not include any inappropriate or offensive words, we translate them into French using DeepL and include them in a dedicated database. It is to be

noted that during all of our experiments, we did not see any offensive output (see more details in Section 4.6).

The agent is then connected to this database containing the different educational resources and the corresponding cues, for the three conditions. The cues for the hand-crafted agent are included manually in the database. Depending on the child's assigned condition, the agent composes the dialogue utterances in order to include the appropriate support. We changed the utterances between the questions to avoid repetition in the agent's dialogue: **Example :** If, for the first question, the agent says **"Here are two cues to help you think of a divergent question about the text, can you use them ?"** than for the following one it will choose another replica such as **"Can you combine these two cues to generate a divergent question about the text ?"**.

4.3 ETHICAL CONSIDERATIONS

One of the first challenges we faced during the design of our LLM-based system is ensuring safe interactions with children. Indeed, despite the impressive advances, LLMs are still considered as 'black-box' systems that are impossible for us to control and that can generate stereotypes and wrong/ biased content [21, 29]. For this reason, we choose to work with a human-in-the-loop setup aiming to have human evaluators verify the content generated by the LLM before proposing it to children. We also chose to use such an offline configuration (i.e. not having the model connected directly to the web platform) in order to ensure the privacy of children's data and avoid having their personal data retrieved by the LLM.

Another risk we encounter with proposing such LLM-based systems is children becoming excessively reliant on these LLM-based systems. This can indeed affect their problem-solving and creative skills [122], as well as their 'real-life' interactions. To address this, it is very important to accompany the use of such systems in schools with pedagogical interventions that aim to raise awareness about the nature and the limitations of LLMs and how to use in an informed manner [4], both for students and teachers. This can help children develop critical-thinking when faced with LLM-generated content and see that it cannot replace other sources of information like books, interactions with classmates, teachers, etc. It can also help teachers use such tools efficiently without affecting the quality of their teaching (e.g. to avoid long and redundant tasks like writing several exercises, to improve the learning experience and enhance task engagement, etc).

It is finally very important to note that the use of such systems should be monitored by the teacher to respond to their pedagogical needs (i.e. frequency of use, collaborative vs. isolated use, types of interactions allowed, etc).

4.4 EXPERIMENTAL PROCEDURE TO EVALUATE GPT-3-DRIVEN "KIDSASK"

The study was approved by the Inria's ethics committee (certificate n°2019-23) and ensured to collect all participants parents' signed consents that contained details about the study's motivation, expected results, measures, etc.

4.4.1 *Participants*

We recruited 75 4th grade students from three classes in two French primary schools. They were between 9 and 10.5 years old and assigned either to the group 1 with the incentive hand-crafted agent (24 with 13 boys and 11 girls), group 2 with the incentive GPT-3-driven agent (26 with 13 boys and 13 girls) or to group 3 with the open GPT-3 agent (25 with 11 boys and 14 girls). The groups were assigned with a pseudo-randomized method after collecting profile data regarding the age, familiarity with using digital devices, curiosity trait and perception measures and reading fluency.

As shown in Table 3, we had three balanced groups in terms of these initial profile measures, with the exception of the familiarity with using digital tools measure. But since the study was run in a controlled environment where researchers could help participants when they have any technical difficulties, we hypothesize that this difference does not have a significant effect on our measure of interest, i.e. divergent QA behaviors. To validate this assumption, we run an ANCOVA analysis in order to investigate the effect of the participants' exposure to digital tools score on this measure between the three groups. Results showed that the interaction was not significant between the three groups (p-value=0.23) suggesting that the exposure measure did not affect the equilibrium between our participants.

4.4.2 *Procedure*

The study consisted of three sessions that lasted between 45 minutes and 1h15, within the same week: one session for collecting the pre-intervention and participants' profile measures, one for the curiosity training (using the two work-spaces of "KidsAsk": WS1 for self-reflection and WS2 for divergent QA training) and one for the post-intervention assessment. The study timeline is the same described in the Timeline Figure 3.3 in the previous chapter, except that we do not use the third work-space of exploration (Day 3 in the figure). See Figure 4.3 or the new timeline and measures.

Measure	Group 1	Group 2	Group 3	pval
Age	9.4 ± 0.43	9.56 ± 0.38	9.52 ± 0.44	0.31
Digital device use	27 ± 7.66	31.39 ± 6.12	32.57 ± 5.28	0.008*
Curiosity trait	27.2 ± 4.38	29.69 ± 4.53	28.36 ± 4.75	0.16
Curiosity perception	37.54 ± 6.72	38.3 ± 6.66	33.92 ± 9.12	0.09
Reading fluency	279 ± 93	297 ± 104	293.2 ± 73.6	0.8
Div. QA fluency	1.33 ± 0.8	2 ± 1.17	2.2 ± 0.93	0.11

Table 3: Profile measures for the three experimental conditions

SESSION 1 During the first session, we presented the study to participants and collected general measures : age, gender, exposure to digital tools and curiosity trait data [144]. We also have the reading fluency using the standardized test in [138] and divergent QA fluency test using an offline task.

Items for the curiosity trait and exposure to digital tools questionnaires can be found in Appendices A.1 and A.7. Procedures for the reading fluency and offline divergent QA tests are also available in AppendicesA.9 and B.2. Details about all measures we use are also available in the upcoming section.

Finally, we also took time to present the study and its aim and explained what divergent questioning means and highlighted the difference between divergent and convergent questions.

SESSION 2 Similar to our first study with "KidsAsk", participants started the second session with the WS1 of the app, dedicated for encouraging self-reflection through a 18-item quiz. For each question, participants could skip it or else, report the confidence in their answers. This is identical to the WS1 presented in the previous study (see Figure 3.4).

The procedure in WS2 for the divergent QA training is also similar to what we presented in the previous study. Participants worked on the exact same texts we presented in the first version of "KidsAsk". Once they read a text, they start an interaction with an agent to help them generate divergent questions about it. As explained above in the "Experimental conditions" section, the agent will do this by giving linguistic cues (i.e. questioning words) + a specific type of semantic cues, according to the participant's assigned condition. The two incentive agents (hand-crafted and GPT-3-driven) had the same behavior; their only difference lied in the generation method of the cues presented (by hand or GPT-3). While the "open" had different utterances to go

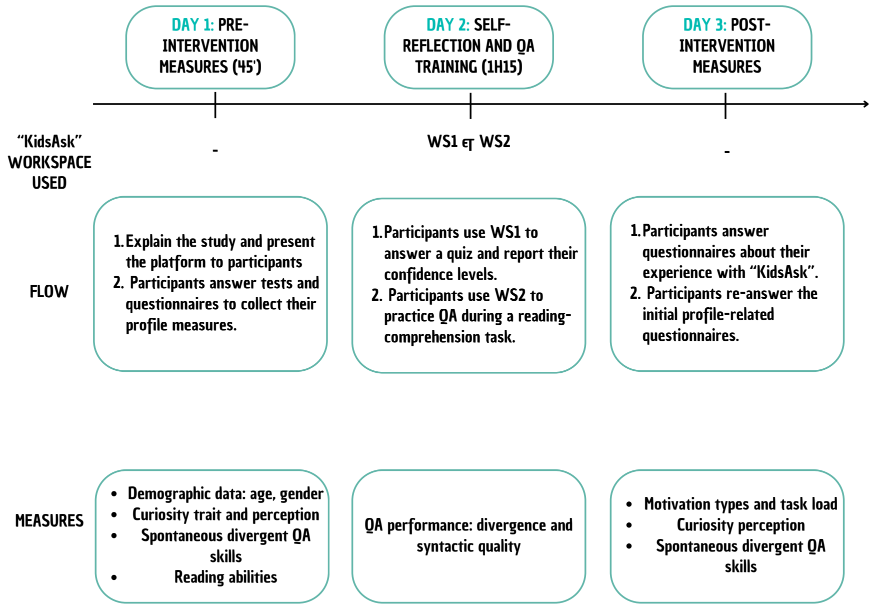


Figure 4.3: Timeline and measures to evaluate the different implementations of "KidsAsk"

with the type of cues it proposes. See Figures 4.4 and 4.5 for the difference between the conditions.

Over the session, children had to generate a total of 18 questions for three different text, 6 questions per text.

SESSION 3 Participants answered post-intervention surveys for the general motivation, types of motivation and the task load. They also re-took the perception of curiosity questionnaire and the QA fluency test (see measures timeline in Figure 4.3 and further descriptions in Section 4.5).

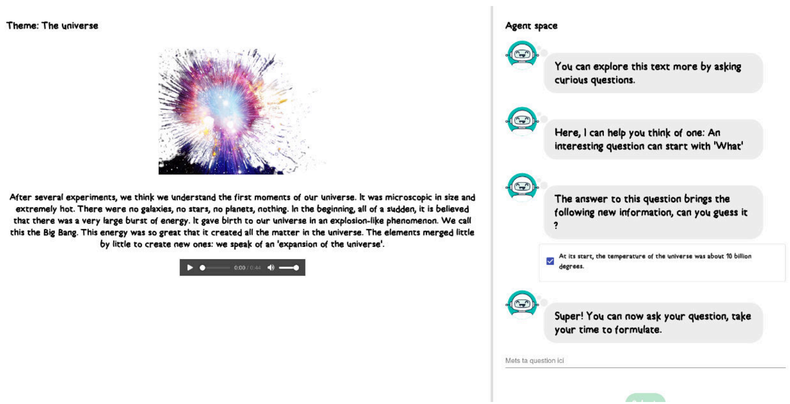


Figure 4.4: Agent behavior for the incentive conditions

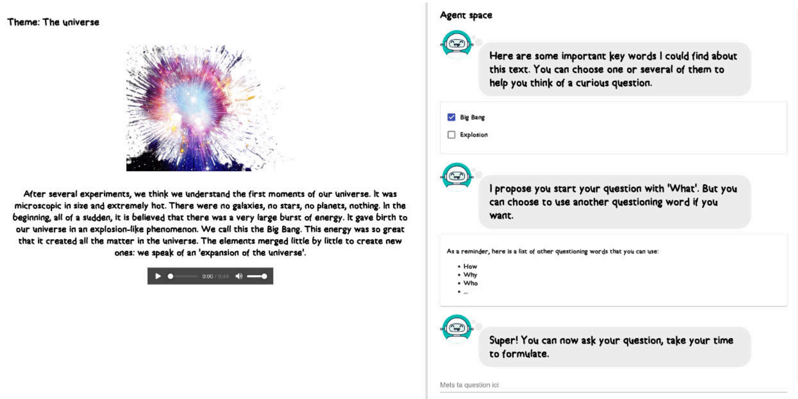


Figure 4.5: Agent behavior for the "open" condition

4.5 MEASURES

4.5.1 *For evaluating GPT-3's performance*

In evaluating the cues generated by GPT-3, two of the researchers that conducted this study perform manual annotations, using specific grids for the linguistic and semantic cues, like the following:

4.5.1.1 *Quality of the linguistic cues*

In studying the quality of the linguistic cues (i.e. the questioning words), we inspect two dimensions and compare them to the hand-generated ones: 1) the variance of the words proposed, i.e. to which extent the agent proposed different words for the questions to be generated. Concretely, this refers to assessing the distribution of the questioning words proposed. And 2) their complexity level, i.e. whether they are simple or compound words.

The complexity level is an important dimension to evaluate. Indeed, according to studies such as [103], the complexity level of the questioning words can predict the divergence level of the question. For example, words like "What difference", "What if", etc are more likely to lead towards divergent questions than simple words like "What", "Where" etc. We therefore decide to compare the proportions of occurrence of this type of compound questioning words when generated by hand vs. by GPT-3.

Details about the procedures followed to perform the annotation of the linguistic skills can be found in Appendix B.7.

4.5.1.2 *Quality of the semantic cues*

In scoring the quality of the semantic cues, we also use manual annotations and compare the scores of the content generated manually vs. with GPT-3. We consider two dimensions: 1) the semantic relatedness to the educational resource in question and 2) the divergence level of the cues, with respect to the resource.

Semantic relatedness is an important measure given that the goal is to lead children to ask divergent questions that are still related and relevant to the educational resource in question. For this we use a 4-point Likert to score a question's semantic relatedness. 0 point corresponds to cues that do not at relate to the text's context (i.e. does not relate to the general topic of the text) and 3 points for cues that are tightly relevant, not only to the text's general topic but more specifically, to the key ideas/ events discussed in it.

We also measure the divergence level of the cues with respect to the associated educational context. This is an important dimension for our work as it helps understand whether the support we're giving participants is pushing them towards divergent QA in the first place. We also use human annotation to assess this dimension, using a 3-point scale. It is to be noted that we only apply this for the incentive cues (i.e. the short sentences that represent knowledge gaps) since the 'open' one are separate words making the classification of the divergence level not relevant.

Details and examples of the annotation procedure for the semantic cues are available in appendix B.8.

4.5.2 *For evaluating participants' performance*

We collect two measures to evaluate children's divergent asking skills: one during the training with the help of the agent and one during an offline test with a different learning activity and where we present no external help or incentive to formulate the questions.

4.5.2.1 *During the training*

We evaluate participants' divergent QA performance exactly as we do in the previous study with "KidsAsk": we retrieve all questions generated during the training session in WS2, we apply specific criteria to filter which questions can be considered for the analysis, we perform a manual annotation to see which of these 'accepted' questions are divergent and, finally, we calculate the percentage of divergent questions generated during the session.

We also assess the syntactic quality of the questions generated using manual annotations following a standardized grid. The grid scores a question from 1 to 8 and has components to assess: whether the question is high- or low-level, its syntactic construction and the use of the questioning word (simple, complex, in-situ or nonexistent).

Details about these two manual annotation procedures can be found in Appendices B.2 and B.4.

4.5.2.2 *During offline tests*

In order to investigate the impact of our training on children's spontaneous divergent QA skills, when they have no external help to formulate their questions, we use an offline test that we administer before and after the training and investigate the change in time.

The test consists of measuring the number of divergent questions that children can generate spontaneously and with no help within 2 minutes, after reading a short text offline. Details about criteria to accept and judge the divergence quality of questions are available in Appendix B.2.

4.5.3 *For evaluating the learning experience*

We use two measures to assess children's learning experience using our platform: 1) Vallerand's scale to distinguish between intrinsic, extrinsic and a-motivation, developed in [223]. And 2), the task load using the Nasa-Tlx questionnaire [92].

Items for both questionnaires are available in Appendices A.5 and A.6.

4.6 RESULTS

Our results section will have three big parts: 1) evaluation of GPT-3-generated cues in terms of general acceptability, syntactic and semantic quality. 2) the impact of the training on children's divergent QA skills and the difference between the three conditions. And 3) the participants' learning experience for the three groups.

4.6.0.1 *GPT-3's performance in generating the cues*

We start with assessing GPT-3's performance in generating the cues. We focus on three metrics: general acceptability (offensiveness, relatedness), linguistic quality (complexity and variety of the questioning words proposed) and semantic quality (divergence level, semantic relatedness).

4.6.0.2 *General aspects*

Before adding the cues generated by GPT-3 to the agent's database, we start by verifying the general quality, in particular with respect to offensiveness (e.g. containing violent, harmful language, etc) and general relatedness to the associated pedagogical resources.

For the offensiveness measure we run a human annotation with a 5-Likert scale (from "Not at all offensive for a 10-year-old child" to "very offensive for a 10-year-old child"). Results showed that 100% of the data generated was scored as no at all offensive (m=1 ; SD =0). We then evaluated the generated cues depending on their relevance to

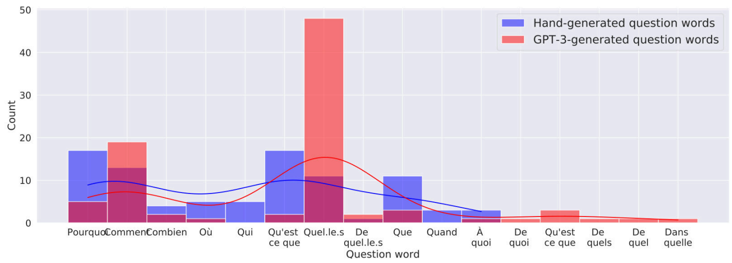


Figure 4.6: Generating the linguistic cues: GPT-3- and hand-generated cues had similar distributions.

the general context of the text (e.g. for a text about the Big Bang, datapoints about the universe in general are considered relevant, however output about different topics like the sports, etc are considered irrelevant). Here again, we run a human annotation and saw that 100% of our generated cues was annotated as relevant to the task's context.

4.6.0.3 *Linguistic quality*

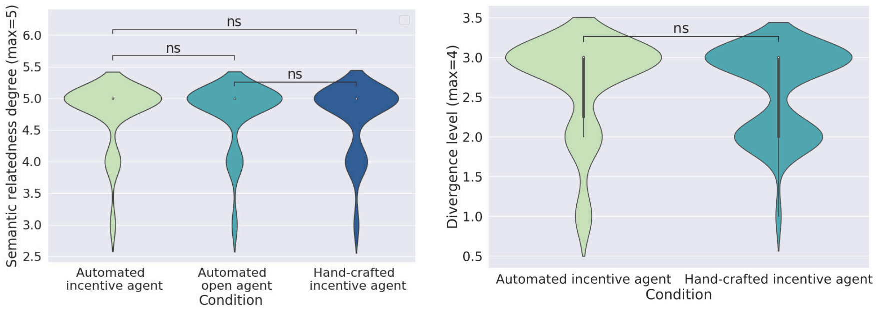
As mentioned above, we evaluate the quality of the linguistic cues by annotating them following the grid presented in the section above (Section 4.5). We assess the variety and the complexity level of these cues and then proceed to compare these scores to those generated by hand.

To assess this, we plot the histogram of the cues generated by hand and by GPT-3 and investigate the similarity between the variances of the two distributions (See Figure 4.6). We use a Levene's test given that the question words generated by hand deviated from normality. Results showed no support to reject similarity between the variances across the two conditions (p-value= 0.65).

We then studied the complexity levels of these cues by comparing the proportions of occurrence of compound words when generated by hand vs. by GPT-3. Again, the 2-sample z-test showed no support to reject the hypothesis of having two similar proportions (p-value = 0.51).

It is to be noted that we only compared the two incentive groups (group 1 and 2) for these dimensions. Indeed, it didn't make sense to compare with the group 3 also since these latter had a slightly different behavior for this cue: given that the open agent cues can lead to several different possible questions, the agent presented participants with a list of standard questioning words that did not vary from one trial to the other, these words could be used for different questions (see Figure 4.5).

With these results, we see similar behaviors between human experts and GPT-3 in generating appropriate questioning words that can lead to divergent QA behaviors.



(a) No difference is seen between the semantic relatedness scores of the cues generated by GPT-3 (both incentive and open) and by hand.

(b) No difference is seen between the divergence level of the semantic incentive cues generated by hand or by GPT-3.

Figure 4.7: Generating the semantic cues : Evaluation with respect to the semantic relatedness to the context and the divergence level.

4.6.0.4 *Semantic quality*

Here, we base our evaluation on comparing two dimensions of the semantic cues generated by GPT-3 for the incentive and the open agents to those generated by hand. We start by investigating the semantic relatedness degree of the cue to the context of the proposed educational resources, as described in Section 4.5. As seen Figure4.7, we see no significant difference between the three conditions for this measure. The one-way ANOVA test failed to exclude similarities between at least two groups (F(2,74)=0.41, p-value = 0.66), with a power of 0.98. Assumptions about the normality of our distributions, homogeneity of variances and independence between the observations were confirmed pre-running the test.

We also investigated the divergence level of the proposed cues for the two incentive agents. This is an important dimension to be assessed given that these very specific cues will heavily condition children's QA performance (i.e. the more they relate to a divergent idea about the text, the more they can lead participants to ask divergent questions). However, this is not the case for the open agent's cues, i.e., the keywords, as the questions' divergence levels do not rely on the words themselves but on how children choose to use them to find a question they're interested in. For this reason, we only assess the divergence level of the cues for the two incentive agents, by running a human annotation following the grid described in Section 4.5. As it can be seen in Figure 4.7, the t-test showed that the two distributions did not differ (t=0.9, p-value=0.37 and power=0.14).

Similarly to the linguistic cues, our results here suggest that our approaches were successful in generating semantic cues that are closely related to the context in question and with high levels of divergence, compared to what we obtained when we had experts manually generate such cues.

4.6.1 *Participants' divergent QA performance*

Moving on to studying participants' interaction with the system, we start by looking into their QA performance during the training, across the different experimental conditions. We then investigate the relationship of this performance with the curiosity trait scores and evaluate the learning experience in terms of its pleasantness and its perceived cognitive load. We finally report the mid-term effect of our training on children's ability to generate divergent questions when faced with a different context.

4.6.1.1 *During the "KidsAsk" training*

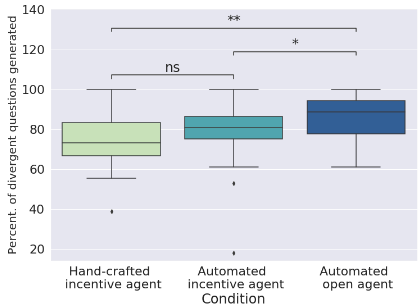
In a first step, we evaluate the general usefulness and accessibility of the cues proposed by the agents in the three conditions by measuring how often children actually choose to use them to generate their questions. Results showed no difference between the two incentive agents in the frequency pf use: M=77.28% and SD=23.08 for the hand-crafted agent; M=76.12% and SD=23.88 for the automated agent with a p-value for the T-test of 0.86 and a power test of 0.05. However, children with the open automated condition (Group 3) ended up using the agent's cues more often (M=91.78%, SD=8.83) and were significantly different from the hand-crafted incentive agent (p-value= 0.03, power=0.85) and the GPT-3-driven one (p-value=0.005 and power=0.81); suggesting that the keywords-based cues may have been easier for children to use.

Moving on, we analyze children's QA performances by comparing the percentage of divergent questions generated for the three experimental conditions, using the manual annotation procedure explained in Section 4.5).

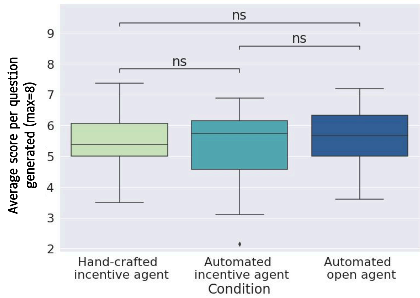
We perform a one-way ANOVA test between the three groups. Results show indeed a significant difference in the performances (F(2,72) =4.11, p-value=0.02). We then perform pairwise T tests: we see no significant difference between the two incentive agents, i.e. the hand-crafted one and the GPT-3-driven one (t=0.88, p-value=0.38 and power =0.14). However, the group that interacted with the automated open agent had a significantly better performance than the Hand-crafted incentive agent (t=-3.17, p-value=0.003 and power=0.82) and the GPT-3-driven incentive agent (t=-1.88, p=0.04 and power=0.45); see Sub-Figure (a) of Figure 4.8.

The normality and the homogeneity of variances of our data were verified pre-running our tests.

We also investigated the quality of the questions generated, following the grid described in Section 4.5. Results showed no signifi-



(a) Participants with the incentive agents had similar divergent QA performances. Those who had the automated open agent had a significantly better performance.



(b) The average quality score per question generated was similar among all participants.

Figure 4.8: Question-Asking performances during the training using the WS2 of "KidsAsk"

cant difference between the two incentive agents (t=0.57, p-value=0.57 and power=0.08). The same observation was also found between two automated agents, i.e. incentive vs open (t=-0.81, p-value=0.41 and power=
0.12) and between the hand-crafted incentive agent and the the automated open one: t=-0.29, p-value=0.77 and power=0.06. See Sub-Figure (b) of Figure4.8.

Taken together, these results show that participants who interacted with the incentive hand-crafted CA had an overall similar divergent QA performance compared to those who had the GPT-3-driven one, suggesting the validity of the GPT-3 prompting approach we used to generate the incentive, 'closed' cues. Furthermore, they suggest that using GPT-3 to generated more 'open' cues, i.e. keywords, may be a stronger strategy to help stimulate children's divergent QA behaviors. Indeed, this interaction helped children formulate more divergent questions while maintaining similar syntactic scores even though they had no semantic formulations for their cues.

Finally, we investigated the interaction between the participants' divergent QA and their curiosity trait (as reported by their parents using the questionnaire in Appendix A.1). To do so, we performed an ANCOVA test between the three groups, with the percentage of divergent questions generated during the training as a dependent variable and the curiosity trait score as a co-variate. Results indicate a significant interaction between the divergent-QA performance and the curiosity scores, within the three conditions (F(2,71) = 4.06, p-value=0.02). Pairwise comparisons are then conducted in order to identify groups that are statistically different : we apply Bonferroni's multiple test correction. The post-hoc analysis show a statistically-significant difference only for the automated open agent condition

Figure 4.9: Participants' curiosity trait scores were strong predictors for their divergent QA performance, but only for those who interacted with the automated open agent.

(p-value=0.006). We also tested the significance of differences between the Pearson correlation coefficients for the three groups, by applying a Ronald Fisher z-transformation.

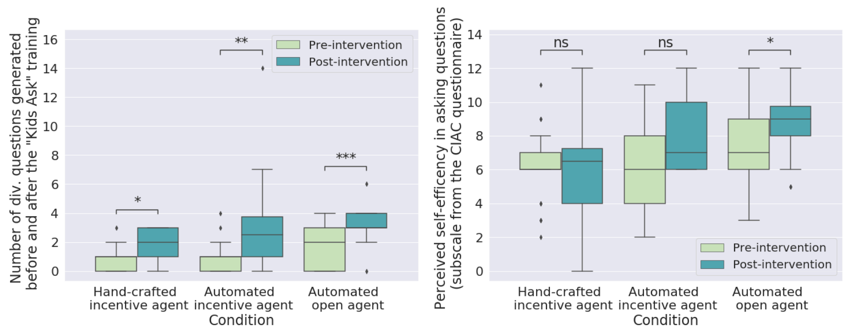
As seen in Figure 4.9, correlations were similar for the two incentive agents (z=1.7, p-value=0.14); however, the open agent led to a significantly different relationship between the hand-crafted incentive agent and the automated open one: z=1.7, p-value=0.04 as well as between the automated incentive agent and the automated open one: z=2.8; p-value=0.002.

These results may suggest that our open agent is more associated with curious thinking given that it leaves more space to children to formulate questions of their own choice and thus, to express their own curiosity. However, in the case of the two incentive agents, we saw no significant correlations between their performance in the task and their curiosity trait. This can be explained by the idea that, given that children were restrained to think of predefined specific questions, their task was more of a social curiosity training rather than an idiosyncratically one, as explained above in Section 4.2.1.

4.6.1.2 *During the offline test*

We finally were interested in studying whether there was a short-term effect of our intervention, i.e. if the training helped participants be better at asking divergent questions when they are put in an educational context other than "KidsAsk" and have no external help to formulate their questions.

For this reason, we assess the participants' spontaneous divergent QA behaviors, before and after the intervention, using the offline test described in Section 4.5. Once we calculated the scores for this test, before and after "KidsAsk", we performed a two-way repeated measures ANOVA test in order to investigate the impact of time on this score, across the three conditions. Results showed a statistically significant effect of the time ($F(2,72)=46.95$, p-value<0.001) and the condition



(a) Participants from the three conditions were able to improve their divergent QA abilities similarly after the "KidsAsk" training.

(b) Children's perception of their QA efficacy was more enhanced with the training for those who interacted with the automated agents.

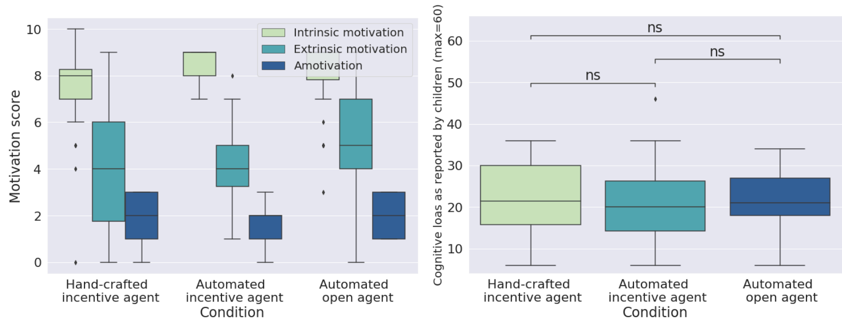
Figure 4.10: Intervention effect : pre- and post-training measures

(F(2,72)=3.9, p-value=0.002), with no effect of the interaction condition:time (F(2,72)=2.23, p=0.11). See Sub-figure (a) in Figure 4.10.

Similarly, we investigated the effect of our intervention on participants' general attitude towards epistemic curiosity. As mentioned before, this is an important measure for us as one of the brakes that can keep children from asking questions is the classroom could be their negative perception of curiosity and their fear of negative judgement [69]. We therefore run a two-way mixed ANOVA test to investigate the difference in the CIAC scores (a validated questionnaire to capture this measure [183]) before and after our intervention, between the different conditions. Results show a non-significant interaction between time and the condition type (F(2,72)=3.87 , p-value=0.06) but reveal a significant change in the scores for all participants pre- and post-intervention (time Effect: F(1,72)=15.74, p-value<0.0001).

We then investigated the Self-Efficiency sub-scale of the CIAC questionnaire which measures how children perceive their own skills in asking relevant and complex questions. As shown the Sub-Figure (b) in Figure 4.10, the three groups had better scores post-intervention and the interaction between time and the groups was significant (F(2, 72)=0.038, p-value=0.04). The post-hoc pairwise tests showed a significant increase in this score only for the group that interacted with the open GPT-3 driven agent (p=0.014). This result aligns well with what's discussed in the design rationale section 4.2.1 where we suggest that giving children tasks with a larger margin of choice and autonomy can effect their perception of their self-efficiency.

Together, these results show that participants from the different groups benefited from the intervention both in terms of divergent QA skills and general perception of curiosity. This suggests the general validity of the training approaches used and the benefits of using CAs to facilitate children's divergent QA skills in general. Also, the



(a) Participants for the the three groups showed significantly more intrinsic motivation than a-motivation or extrinsic motivation.

(b) The workload of the task was similar for the three different conditions.

Figure 4.11: Learner experience measures

open agent resulted in more self-efficiency in asking questions and correlations with curiosity traits, making this latter's strategy to train curiosity rather more relevant.

4.6.2 *Participants' learning experience measures*

Using the Vallerand's motivational scales we describe in Section 4.5 ([223]), we see no difference between the children's intrinsic motivation during the training with the three different agents. However, we saw that the three groups were significantly more intrinsically than extrinsically motivated to do the task (t=3.3, p<0.001, power=0.98 for the hand-crafted incentive agent, t=12.02, p<0.001, power=1 for the automated incentive agent and t=3.9, p<0.001, power=0.97 for the automated open agent); see sub-figure (a) in Figure4.11.

We also investigated the task load as perceived by children, following the Nasa-tlx scale as described in Section 4.5 ([92]). The aim here is twofold: 1) evaluate whether or not children perceived that the answers provided by GPT-3 were harder to assess than those provided by humans. And 2) see if the lack of a semantic formulation in the open cues for the open automated agent makes the question-generation task harder for children. To address these goals, we perform pairwise t-tests. Results showed no significant differences between the two incentive agents (t=-0.4, p=0.1, power=0.06) or between the incentive and open agent (t=-0.3, p-value=0.7, power=0.06).

These results suggest indeed the accessibility of GPT-3's output, even when no semantic formulation was given (i.e. only keywords).

4.7 DISCUSSION

With this work, we study using prompt-based learning with the pre-trained LLM GPT-3 to generate the pedagogical content for a divergent QA training. We investigate the validity of our approach by using human annotations to score this content and compare its quality, as well as its impact on children's divergent QA skills, to the hand-generated one.

The proposed prompt-based method showed rather positive results in terms of syntactic quality, semantic quality, and impact on children's divergent QA behaviors. Moreover, we saw that the 'open' cues (i.e. cues that lead to several possible questions rather than a specific predefined one) were more efficient in helping students express their own curiosity using divergent questions. These observations were rather expected given that the open cues were generated with the intention to leave students with more control to choose their own questions during the training. Indeed, we hypothesize that such a setting can lead children to work on questions that correspond to their own competency level which will result in better perception of the self-efficiency in the task [159] and therefore, in more engagement and learning progress [177, 218].

Furthermore, we also investigated the link curiosity as trait (such as reported by parents using the questionnaire developed in [145]) and the divergent QA behaviors. Our results showed a strong correlation only for the group that had the open cues. This result meets our goals of designing the open agent: giving children QA activities that they can transpose more to their zones of proximal development (ZPD) is likely to help their curiosity. It also validates our suspicion that our incentive agents are more efficient on training the linguistic part of the process of generating divergent questions, rather than the curiosity one. Indeed, generating divergent questions requires two main steps: 1) using one's curiosity to identify a knowledge gap and 2) using one's linguistic skills to pursue this knowledge gap by transforming it into a question. And since the first step is led by the agent in the incentive case (either hand-crafted or GPT-3-driven), children are not given a direct opportunity to choose their questions using their own curiosity, like it is the case for the open agent. This is an interesting finding as it positions the assessment of divergent QA behaviors, when they are mostly directed by students themselves, as a behavioral measure for curiosity that is more suited to e-learning environments than the classic self-report measures.

Finally, regarding the motivation scores, we saw high yet similar scores across the three conditions. This was counter our expectations as we hypothesized to see more intrinsic motivation for the open agent group where the task offers more degrees of choice and autonomy. This result could maybe explained by the novelty aspect of our general approach (i.e. using tablets in the classroom, interacting with a CA, the QA task itself, ..). And also, given the short length of the training. Indeed, we can speculate that the motivation scores for the incentive conditions where children have little control over the task, would drop if we had a longer training.

Taken together, our results suggest that the approach taken to facilitate divergent QA was efficient in general, but was more associated with curiosity trait, transfer effects and self-efficiency perceptions for the 'open' condition. Thus, we can propose this new method as a promising approach to train children's curiosity in general and their divergent QA skills in particular. It also allows us to think of new possible implementations that can facilitate the divergent QA task by guiding children to identify the key concepts/words of a resource first, then encouraging them to use these words to identify specific knowledge gaps and formulate a related divergent question. Indeed, we can imagine that decomposing the task in this way can make it easier for children to learn and to re-use in the future. Such an approach can be either adopted by teachers in classrooms or as an instructional design principle to implement in educational technologies.

4.8 LIMITATIONS AND FUTURE DIRECTIONS

One of the main drawbacks to our current implementation of "KidsAsk" is the lack of feedback we give children about the questions they generate during the training.

Indeed, and even though we see enhanced perceptions of the self-efficiency, we do not know if this is directly connected to their perception of their performance or to other factors, such as repeatability of the exercise, bias in the self-report measure, etc. Therefore, one future direction for this work will be to explore ways to use LLMs in general in order to analyze children's questions and give them real-time feedback about their relevance, their divergence level and their syntactic construction. Recent work such in [229, 231] has explored using prompt-learning methods to evaluate the generation of convergent questions and showed rather encouraging results. Therefore, one possible track can be to take inspiration from their methods and explore adapting them to divergent questions.

Another factor to be considered while evaluating this training is its short duration in time. Indeed, children were only asked to do this task during one short session which may have facilitated their engagement and played a role in the positive results we saw. Therefore, one future direction for this work is to propose a longer version of our training and investigate its long-term impact on children's curiosity-driven behaviors.

On another hand, even though prompt-based learning with GPT-3 is quite an easy method to implement, it does still require some knowledge about prompt-engineering in order to generate relevant output. Indeed, we hypothesize that we won't see the same results if we have used less precise and concise prompts and different settings. This leads us to think about the different trainings we should provide teachers with in order to ensure good practices and satisfying results. Moreover, and as mentioned earlier in the "Ethical considerations" section, our implementation raises some educational challenges: having children use LLM-powered systems in general without knowledge about their best practices and without developing a sense of critical thinking when interacting with them can be challenging [122, 124]. For example, it can reduce their creativity, engagement when resolving tasks, etc; making the system loose its potential benefits. In this direction, we are planning to develop pedagogical interventions in order to train children's intellectual humility and critical thinking when using LLM-powered educational tools [4].

4.9 CONCLUSION

This work is a part of a larger project attempting to develop technology-enhanced approaches to train curiosity in children. The study proposes to address some of the brakes that keep them from seeking new information autonomously such as linguistic barriers and negative perceptions of curiosity. More particularly, we carefully investigate the potential of LLMs to generate pedagogical content and propose a novel GPT-3-driven system that trains divergent QA using simple and efficient prompt-based methods.

Our results, including a field study, showed the benefits of using such tools to engage children in divergent QA behaviors and improve their curiosity perceptions. This work motivates the implementation of such curiosity-eliciting approaches both in standard classroom settings and in e-learning environments using NLP methods.

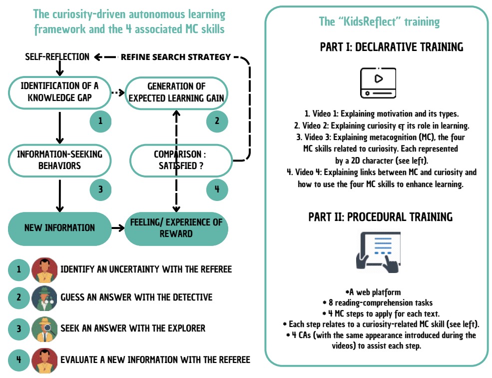
5 STUDY 3. DESIGNING, IMPLEMENTING AND EVALUATING "KIDSREFLECT": A TRAINING FOR CURIOSITY PERCEPTIONS AND RELATED METACOGNITIVE SKILLS

Collaborators: Edith Law, HCI Lab, University of Waterloo, Canada.
Chloé Desvaux, University of Bordeaux.

Aims: This chapter aims to introduce "KidsReflect", a training for children's perceptions of curiosity and their curiosity-related metacognitive (MC) skills.

Abstract: This chapter outlines the rationale behind the design of "KidsReflect". It presents two versions: one that uses animated videos to give declarative knowledge about curiosity and four specific MC skills we hypothesize are crucial to support it. And one that also adds a procedural training using a platform with CAs that help practice these skills during a reading-comprehension task by proposing specific cues. The chapter then details the setup used to evaluate the training: 86 students between 9 and 11 years old had either: the declarative + procedural "KidsReflect", the declarative only, or no "KidsReflect". Then, they all had the QA training with "KidsAsk". Finally, the chapter presents the results: the declarative + procedural intervention had the strongest effect on students' metacognitive sensitivity, perception of curiosity and divergent QA performance during the online "KidsAsk" training. The increase of these three factors, along with the curiosity trait, predicted the post-intervention spontaneous divergent QA behaviors.

Contents



Overview of the "KidsReflect" training

A pilot study of the work in this chapter is published in IDC'23: ***Rania Abdelghani et al. "Interactive environments for training children's curiosity through the practice of metacognitive skills: a pilot study."*** *In:* **Proceedings of the 22nd Annual ACM Interaction Design and Children Conference.** ***2023, pp. 495–501****. A journal submission is currently being prepared.*

5.1 introduction

During the two studies presented in the previous sections, our main focus was to develop and test an intervention for training divergent question-asking; a skill we suggest has a tight correlation with epistemic curiosity [8]. Our approach was based on having a CA to guide children into seeing their knowledge gaps (either by proposing an explicit and specific KG or a series of key words that can be used to formulate their own KG). The agent then helps children express these KG in the form of a divergent question by offering some linguistic help.

While we observe positive impacts on children's spontaneous divergent QA behaviors after such trainings, our strategies may prove insufficient for fostering the development of curiosity and divergent thinking processes on the longer term. Indeed, it lacks giving opportunity for children to discover their own learning goals independently, which is a crucial step in these processes [196]. This is highly important as it means that our strategy may be overlooking the "identification of a knowledge gap" part— an integral component inseparable from the curiosity-driven learning process [147]. Furthermore, our training failed in enhancing children's social perceptions of curiosity and QA behaviors, which at its turn could be a significant brake to their adoption of such behaviors [69].

In this context, we therefore make the hypothesis that, in order to reinforce the impact of our training on divergent QA and curiosity, two distinct tracks can be explored:

1. correct children's misconceptions of curiosity and question-asking It is important to extend the training to introduce and explain curiosity and QA behaviors to children: their function, their benefits for learning, and how to control and use them efficiently. Indeed, children tend to have about these behaviors (useless, related to stupidity, etc) and thus, fear to adopt them in their classroom [8, 111, 183].

2. train the skills related to knowledge gap identification, monitoring and evaluation Following curiosity models, we want to focus on practicing the faculty of problem discovery and knowledge gap identification, monitoring and evaluation. More concretely, this means to go from a training that makes the KG explicit and then helps children formulate it as a divergent question to one that starts by helping them find their KGs on their own and pursue them and without the need for external help. This is indeed challenging as it requires children to be able to decide which KG to follow in their vast environment, what behaviors to adopt to compensate for them and deciding when satiation is attained. These are all complex behaviors that partly rely on metacognition, i.e. the ability to understand and monitor one's own knowledge and learning strategies [77, 107, 168]. Several studies [226] identify metacognitive skills such as the evaluation of one's current state of knowledge and the prediction of learning progress as necessary conditions for triggering and maintaining curiosity as they can facilitate continuous identification of learning goals through divergent thinking.

Based on these two directions, the current work proposes to bring the two ideas in order to design a new pedagogical intervention that aims to : 1) enhance children's perception of curiosity and QA behaviors, by introducing them to curiosity, its importance and how to leverage it to learn better and enjoy learning. And 2) train the metacognitive skills that are suggested to facilitate the triggering and maintaining of curiosity states. To identify these skills, we take inspiration from the theoretical framework explaining curiosity-driven learning mechanisms developed in [168]. By attempting to link its main components to specific metacognitive skills, we identify four curiosity-related ones: 1) the ability to be aware of missing information in one's knowledge, 2) the ability to estimate the expected learning gain that will result from seeking a specific information, 3) the ability to seek the missing information using the appropriate strategy and, finally, 4) the ability to assess the result of the curiosity-driven process and decide of subsequent actions.

We use these ideas to design a new pedagogical training we call "KidsReflect", aiming to foster children's curiosity-driven learning, through enhancing their social perceptions of curiosity and the practice of the said curiosity-related metacognitive skills. "KidsReflect" had two versions: one that gives declarative knowledge about curiosity and the four related MC skills stated above, using animated videos we created within the team. And one that, beyond the videos, also adds a procedural part to practice these four skills during a reading-comprehension task, using an online platform we designed. Children's goal is to gain new knowledge about the texts independently, using the four curiosity-related MC skills. To practice these skills, they have the help of conversational agents that give them specific hints and suggestions to guide them in the task.

We study the efficiency of the two versions of the training by investigating its impact on four main outcome measures: children's metacognitive sensitivity, their perception of curiosity and QA, their ability to benefit from an online divergent QA training ("KidsAsk") and finally, their offline spontaneous divergent QA behaviors.

5.2 Design rationale of "KidsReflect"

"KidsReflect" contains two dimensions. The first one is designed to give declarative knowledge about curiosity-driven learning and its importance in order to help children understand it better and have more positive perceptions about it. The second component is designed to give both declarative and procedural knowledge about curiosity-related metacognitive skills [62, 128].

The first dimension being rather explicit (i.e contains exclusively instructional content about curiosity, its links with memory, attention, etc), we only present its content in the section below (Section 5.2.2.1). In this current one, we focus more on the approach we led in order to come up with the instructional design principles and the pedagogical content for the second dimension of "KidsReflect", focusing on metacognition.

To develop this dimension, we followed two main steps:

1. Identify the metacognitive abilities that can directly nourish or interfere with curiosity mechanisms. To do this, we start by presenting the autonomous learning model proposed by Murayama et al. [168]. We then try to operationalize this model by proposing explicit articulations between its key components and the metacognitive faculties needed to achieve them.

2. Once identified, we focus on the design of pedagogical activities that can directly target the exercise of these metacognitive skills involved in curiosity-driven learning.

5.2.1 *Linking the curiosity-driven learning process to metacognition*

Looking at this framework proposed by Murayama in [168], we find four relevant articulations between its key components and metacognition. See an illustration of the framework as proposed by the authors and the four links we make with specific metacognitive faculties in Figure 5.1.

These articulations are four-fold:

5.2.1.1 *1. Metacognition facilitates curiosity stimulation through enhancing evaluative self-reflection*

As suggested in the Figure 5.1, the activation of epistemic curiosity mobilizes the ability to identify a knowledge gap (i.e. a discrepancy in one's knowledge) that they want to resolve and that they consider as achievable. This is aligned with the curiosity theories focusing on the information gap theory [147]: these latter distinguish two types of unknowns—known unknowns (knowledge that we recognize as being missing) and unknown unknowns (knowledge that we do not know is missing). The crux of the problem here, therefore, is how to get an individual to transform unknown unknowns into known unknowns in order to motivate curiosity-driven information-search. This is a challenging procedure as it requires high-level skills such

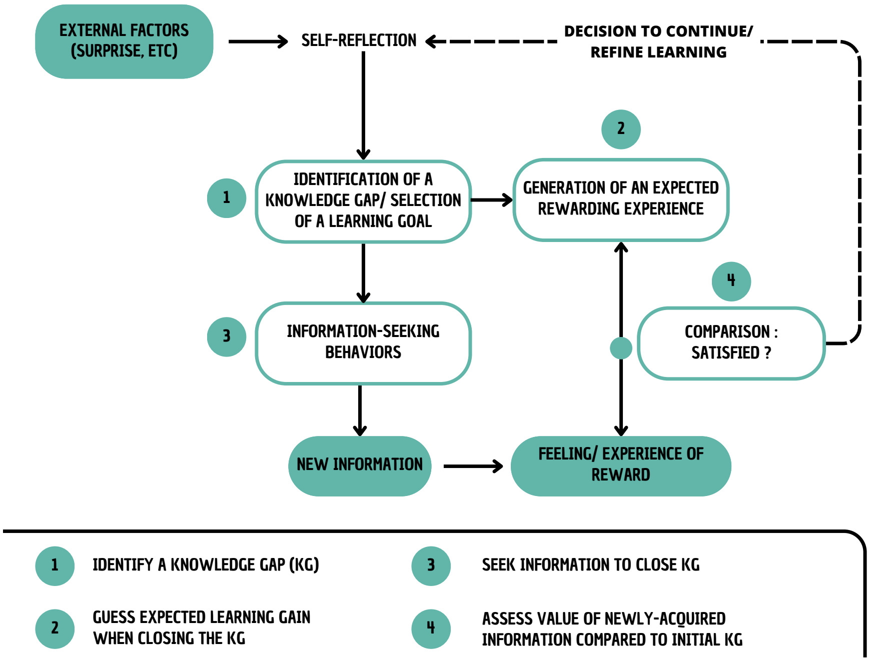


Figure 5.1: Linking metacognitive regulatory and evaluative skills to the autonomous curiosity-driven learning framework as proposed by Murayama et al. in [168]

as linking new stimuli to previous knowledge in order to discover a precise unknown, problem, incongruity, etc. It is also suggested to be in direct link with the metacognitive ability to evaluate one's own knowledge level given a specific learning context [77]. Several studies have proposed that children lack this ability and tend to bypass their uncertainties as they often overestimate their knowledge levels [137, 193].

We therefore posit that the activation of epistemic curiosity requires the ability to identify an uncertainty, or a knowledge deficit (i.e. a problem discovery, a learning goal) using self-reflection, causal attributions, idea-linking, etc. This represents the first MC skill we identify as essential to train in order to foster curiosity-driven learning process; we call it the "IDENTIFY" skill (see Figure 5.1).

5.2.1.2 2. *Metacognition amplifies feelings of curiosity through helping set realistic expectations*

As seen in the Figure 5.1, generating an expected learning gain plays a central role in the curiosity-driven learning process. It seems to act as a mediator in maintaining information-search behaviors: once a knowledge gap is identified, generating an expected knowledge gain (e.g. by formulating a hypothesis about the knowledge gain) and comparing it to the result of the actual information-search can determine whether future curious behaviors will occur [168].

But other than this role (that we will be focusing on during the fourth articulation below), predicting rewarding experiences can also amplify and regulate initial feelings of curiosity [60, 168, 185, 206]. Indeed, it can increase the salience of the identified knowledge gap and make it more attractive to search for [147]: individuals are most curious about stimuli they think they almost have the answer to [82, 85]. Studies such as [32, 184] show indeed that asking individuals to generate hypotheses before seeing the correct answer to a question leads to higher curiosity states. Children also are suggested to evaluate the informational potential to be gained, via predicting reward signals, before engaging in exploratory behaviors [182].

However, generating such expected learning rewards with optimal value (in other words, making informed and educated guesses about a knowledge gap) is not trivial. It requires individuals to use previous knowledge and knowledge about their current cognitive state to project themselves into a stronger one, yet one that still feels comfortable enough for them to be motivated to pursue. This heavily relies on metacognitive faculties as it requires self-reflection and evaluation as well as prior knowledge and weighing of alternatives [220]. Making informed guesses is suggested to be a malleable skill that could be trained via encouraging observation and experimentation [220].

This represents the second MC skill we identify as essential to train in order to foster curiosity-driven learning process; we call it the "GUESS" skill (see Figure 5.1).

5.2.1.3 *3. Metacognition elaborates curiosity through facilitating efficient question-asking*

Seeking information autonomously is an essential part of curiosity-riven learning (Figure 5.1). It can take different forms: exploration, reading books, etc. Here, we focus on question-asking behaviors. Indeed, it is one of the main behaviors that allow children to pursue information and optimize information-search procedures [194].

However, for these behaviors to be efficient, individuals need to adapt them, i.e. decide which specific questions to ask, given the current knowledge status, in order to achieve their learning goal. This means that the ability to monitor one's learning goals, and adopt QA behaviors accordingly is crucial [113]. And although we know that question-asking skills develop from early childhood, children still can fail to dynamically adapt their questions to navigate efficiently through the knowledge space and achieve their goals [161]. Several elements can interfere with this ability, and importantly to this work, metacognition. Indeed, control metacognitive skills are important to help understand the features that make the "informational effectiveness" of a question [194]. Thus, they help identify and control what kinds of questions are needed to be asked to achieve the wanted learning progress.

Based on these ideas, we identify informed question-asking as an important skill to be trained in order to facilitate children's curiosity-driven information-search behaviors. In our framework, we call this the "SEEK" skill.

5.2.1.4 *4. Metacognition maintains curiosity through facilitating self-monitoring*

Maintaining information-search behaviors can heavily rely on the individuals' ability to accurately assess their learning progress (see Figure 5.1). In the framework, asking new questions is mainly the result of comparing the initial learning expectation (i.e. the estimation of the reward expected from closing the identified KG, the generated hypothesis) and the actual learning that was achieved thanks to the information-search behavior (i.e. the answer to the question).

This idea can also be found in other models. For instance, Brydevall et al. suggest that upon the acquisition of new knowledge resulting from a seek process, learners experience a feeling of reward that updates the expected value of future new information and thus, influences the future information-seeking behaviors [35]. A similar idea is also suggested with the Learning Progress Theory (LPT) [176]: learners become most curious when they work on tasks that provide them with an optimal level of learning progress feeling, as perceived and evaluated by the learners themselves. Meaning that the faculty to accurately self-assess and monitor progress plays an important role in maintaining curiosity-driven behaviors. Doing this evaluation of newly-acquired rewards and comparison with expected ones requires continuous metacognitive evaluation of the learning progress. Concretely, this represents the faculty to take time to evaluate newly-acquired information with respect to one's current and future goals and use this comparison to decide of subsequent learning strategies.

We call this MC ability, the "ASSESS" skill in our framework (Figure 5.1), an essential skill to train to help promote sustainable curiosity-driven learning cycles.

5.2.2 *Pedagogical content for the "KidsReflect" training*

As explained above, "KidsReflect" has two main dimensions: one that focuses on introducing curiosity and its role for learning and one that focuses on providing declarative and procedural knowledge about the four metacognitive skills we identified above as correlated to curiosity: IDENTIFY, GUESS, SEEK and ASSESS.

To design this pedagogical intervention, we take inspiration from studies such as in [57] that showed, for instance, that simple declarative workshops about growth mind-set helped students be more motivated and improve their grades. We also looked at classification taxonomies of educational goals in [130] and thus decided to have a two-part training: one that aims to give declarative/ conceptual knowledge about curiosity and related metacognitive skills; and one that aims to help gain procedural/ experience-based knowledge about these latter.

5.2.2.1 *Part I: Declarative knowledge about curiosity and the related metacognitive skills*

This first part consists of 4 sessions. Each session relies on presenting an animated video that explains a concept related to curiosity and/or metacognition as well as their role in learning:

- Video 1: presented motivation in general, its importance for learning and its two types: intrinsic and extrinsic. It highlights the importance of intrinsic motivation and presents curiosity as a special type of this latter.
- Video 2: presented the importance of curiosity in learning, focusing on its positive impact on three main indicators: attention, memory and autonomy/ self-efficiency.
- Video 3: presented metacognition and how to use it during a learning task. It broke up metacognition-based learning into four steps that correspond to the skills identified above: IDENTIFY, GUESS, SEEK and ASSESS. Each skill was presented with an example of use and associated to a 2-D character to facilitate their memorisation.
- Video 4: links curiosity and metacognition by explaining the role of each of the skills shown in video 3 with respect to curiosity-driven learning steps.

See an overview of the content of these videos and links to play them (in French) in Figure 5.2. The story board with the detailed scenario for one of the videos is available in Figure 5.3; the story boards for the rest of the videos are available in Appendix C.1.

As illustrated in Figure 5.2, the third video introduces the characters that represent the four metacognitive skills we identified as directly related to curiosity. In order to facilitate their understanding, we model them with 2D characters that correspond to their role in the curiosity-driven cycle: a first referee to represent the IDENTIFY skill, a detective to refer to the GUESS skill, an explorer to refer to the SEEK skill and a second referee to refer to the ASSESS skill. The idea of modeling MC skills with specific characters is mainly inspired from the "Reflecto" model [68].

The characters adopt specific roles/ strategies when trying to resolve learning tasks (see Figure 5.4):

- **The first referee** : when faced with a new learning task, it reflects on its previous knowledge and connects it to the current

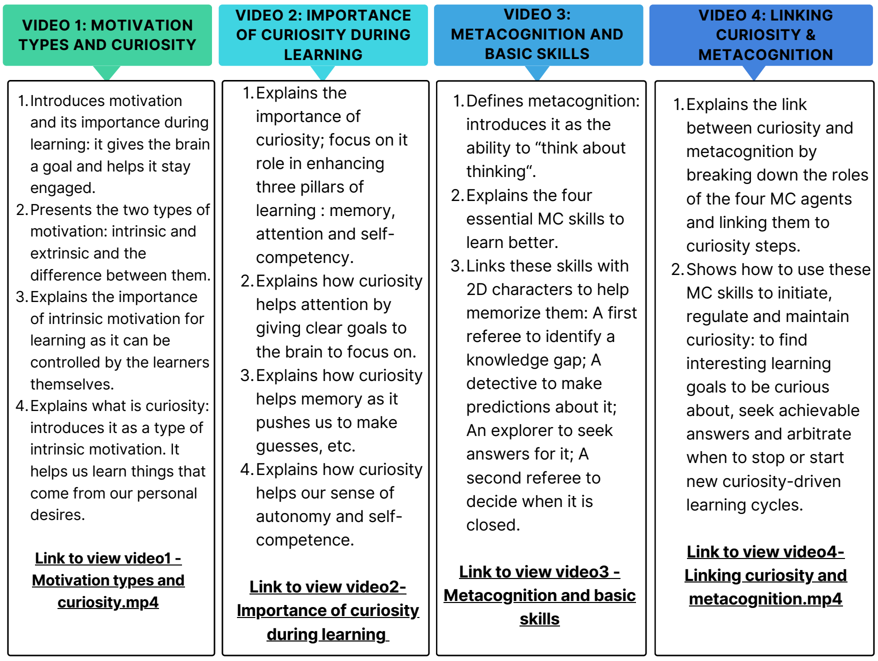


Figure 5.2: Overview of the pedagogical goals and links for the four videos presented during the declarative part of the "KidsReflect" training

context in order to identify novel, surprising, contradictory information, etc. This is then used to identify a specific knowledge deficiency, learning goal or an uncertainty that it wants to pursue. The first referee reflects the IDENTIFY skill.

- **The detective** : once it has a learning goal (given by the first referee), the detective will make educated guesses about this problem in order to further motivate information-search behaviors. It does so by using its previous knowledge, liking ideas, weighing alternatives, etc in order to form a realistic expected learning progress. The detective reflects the GUESS skill.
- **The explorer** : uses information about the learning goal and the expected information gain in order to formulate the relevant and specific uncertainty-driven questions. The explorer reflects the SEEK skill.
- **The second referee** : When receiving answers from external agents to its inquiries, the second referee will evaluate them based on its knowledge about the learning goal and its assessment of the rewarding experience compared to the expected one (generated by the detective). Using these signals, the second referee decides for the efficiency of the search procedure and whether it gives place to subsequent ones. The second referee reflects the ASSESS skill.

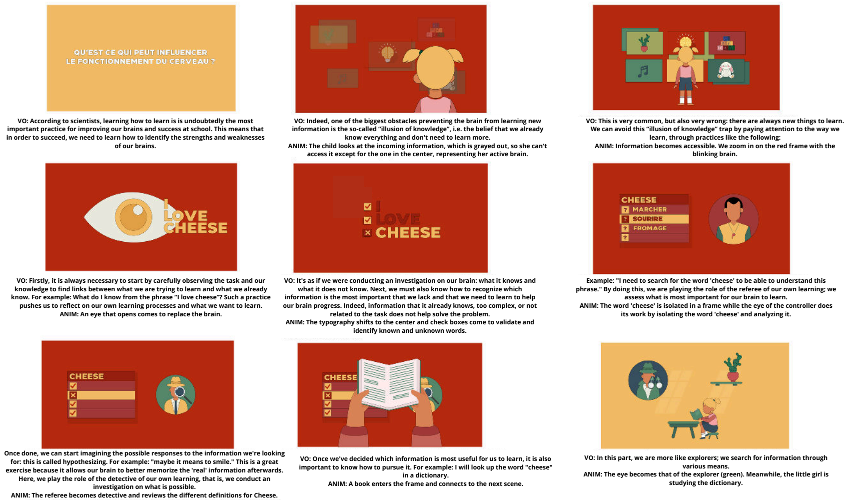


Figure 5.3: Scenario for the third video "Metacognition and basic skills" presented during the training: it introduces children to the metacognition and the four MC skills that are related to curiosity

5.2.2.2 *Part II: Procedural knowledge about the curiosity-related metacognitive skills*

The aim of the second part of the training is to help children put into practice the four metacognitive skills they saw during the videos. For this we create a web-based platform that offers reading-comprehension tasks. Children's ultimate goal is to use the four-MC-step curiosity-driven learning behavior showed during the videos (see Figure 5.3) in order to gain new information about the text at hand in an autonomous and independent way. The platform offers different prompts to help children achieve these four steps successfully.

The prompts are presented to children using a conversational agent that acts like a learning companion. Its goal is to guide them into using the IDENTIFY-GUESS-SEEK-ASSESS chain-of-thought, step by step, in order to gain new knowledge that can deepen their understanding of the text at hand. To do so, and for each new text, the agent is designed to give prompts and suggestions to help use each of the MC skills, one by one and in order (i.e. first, identify a knowledge, then make a guess about it, then seek it and finally, evaluate the result of the search). The agent moved to prompt a specific step only when children achieved the previous one: e.g. the agent moves to talking about the GUESS skill only when the participant enters an identified uncertainty (representing the IDENTIFY skill), the agent moves to talking about the SEEK skill only when the participant enters a guess about its uncertainty (representing the GUESS skill), etc. To help children memorize these skills better, the agent took, for each skill, an appearance identical to the one presented during the videos in the first part of the training: when prompting the IDENTIFY skill, it appeared and was presented as the first referee. When prompting

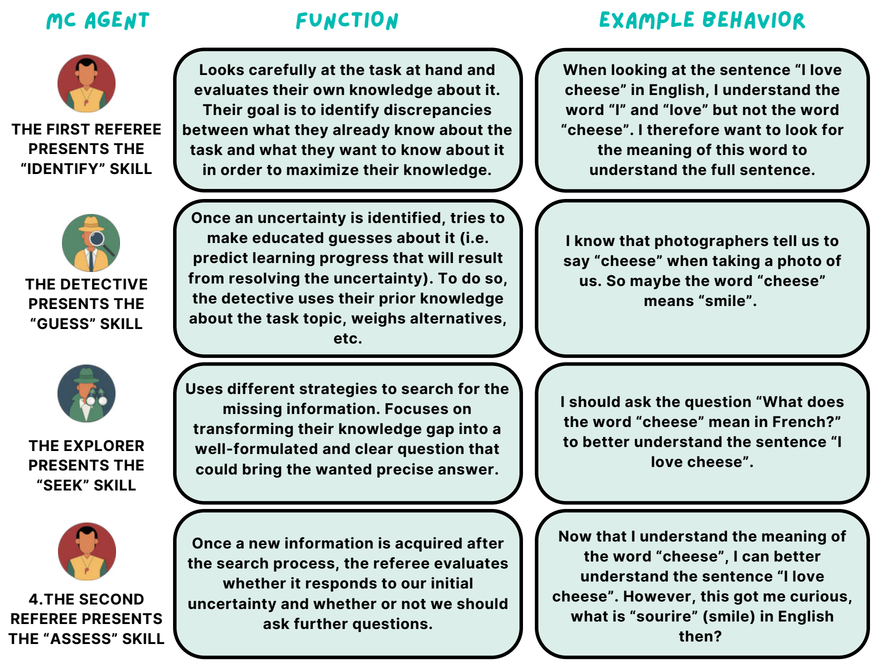


Figure 5.4: Presentation of the MC agents and their functions that children saw during the training. An example of their behavior when resolving a learning task (understanding a sentence is a foreign language) is given.

the GUESS skill, it appeared and was presented as the detective, etc. See Figure 5.5 for an overview of the interaction with the agents when working on a specific text.

Overall, the interaction with the agent is designed to be like the following: participants are first introduced to a short text. They are told that their goal is to gain extended, novel and useful information about it using their curiosity and the four metacognitive steps they saw during the first part of the training (we remind them of these steps in the beginning of the session, before presenting the current task).

We then introduce the platform and tell them that, in order to facilitate their task, they should follow the different agents' instructions, that will appear as the task progresses. Once they finish reading the text, the agent appears as the first referee (representing the IDENTIFY skill) and guides them into identifying and formulating a knowledge gap. To do so, it starts by reminding them of its role as a referee and why it is important. It then encourages them to reflect on their previous knowledge about the text's topic and then to use this to look for information that is contradicting their previous knowledge, they are seeing for the first time, or is surprising for them. It also gives an example on how it does this, similar to a think-aloud approach. **Example for the text in Figure 5.7: "I already know that the Earth has very high temperatures deep down. But I did not know that**

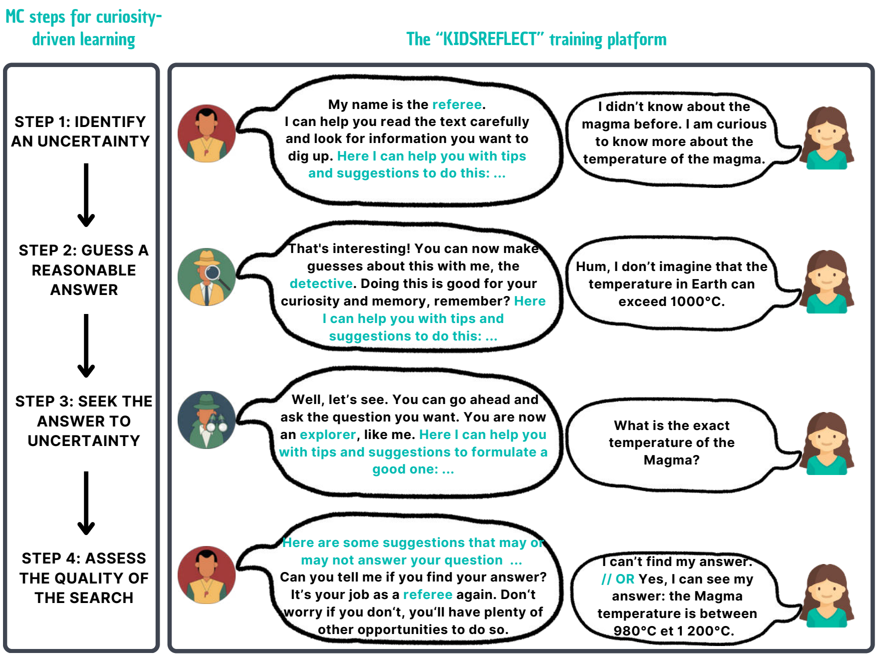


Figure 5.5: Overview of the interaction with the different agents during the reading of one text. Each agent represents one MC skill related to curiosity, it reminds children of its role and gives tips and suggestions on how to use it. It changes roles and appearances each time after children achieve a step.

high temperatures could turn the rocks into liquid. I am thus curious about the temperature that can turn rocks into liquid." See an example of agent's full dialogue in Figure 5.6.

After validating this step (by entering a learning goal/ an uncertainty), the agent changes its appearance and becomes a detective to represent the GUESS skill. Similar to the first step, it starts by reminding its role and why it is important: making good guesses can enhance curiosity and memory. It then guides participant into formulating educated guesses by explaining the properties of the latter: it should be tightly related to the uncertainty, it should rely on previous knowledge, it should not be formulated randomly, etc. Once again, the agent finishes its dialogue by giving an example to present its own approach to formulate an informed guess. **Example: "I know that water tuns into liquid at 100°C. I also know that rocks require higher temperatures to become liquid. So I guess that maybe rocks need more than 100°C to become liquid."** See an example of the agent's full dialogue in Figure 5.6.

Once the prediction submitted, the agent turns into an explorer to represent the SEEK skill. It now helps the participant formulate the appropriate question that could lead to closing their knowledge gap. To do so, it reminds the participant of their goal and gives them examples of generic questioning words they can use to formulate their inquiry. To give a concrete example, the agent also shows children how it changed it own knowledge gap into a question. **Example: "My goal is to know the temperature that could turn rocks into liquid. My question is thus: At what temperature do rocks turn into liquid?".** See an example of the agent's full dialogue in Figure 5.6.

And finally, once the question submitted, we display a set of 3 pieces of information that are related to the text's topic and that may or may not contain the answer to the question asked by the child. The second referee thus appears to lead children into reflecting on these pieces of information and decide whether or not they bring them closer to closing their knowledge gap, or to generating new/ further questions. This agent represents the ASSESS skill. **For example: "I see the answer to my question in the list: rocks can turn into liquid starting from 600°C."** See an example of the agent's full dialogue in Figure 5.6.

The agents had predefined scripted behaviors and did not give any specific feedback regarding children's inputs. It only gave general feedback to congratulate them for achieving a step or completing the learning cycle for a text. Details about this choice are described in Section 5.2.3.

See Figure 5.6 for a concrete example of the agents' dialogue for a given text, Figure 5.7 for a snap of the platform and Section 5.2.3 for details about the technical implementation of the platform.

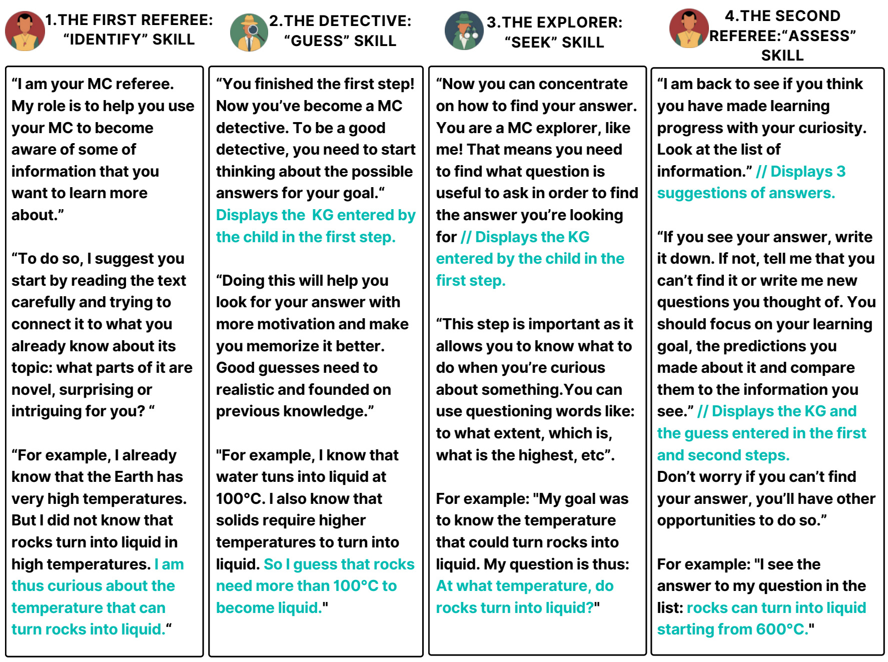


Figure 5.6: Examples of the four agents' utterances for a given text

In total, children had 8 texts, i.e. 8 learning cycles to complete using the four skills. During the first two cycles, and in order to help children get familiar with the task, each agent gave a list of three

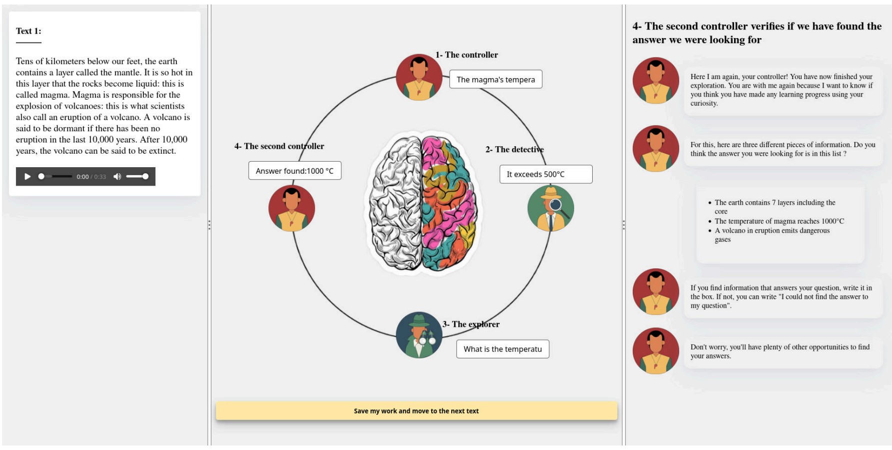


Figure 5.7: Screenshot of the platform used in the "KidsReflect" training during the exercise of the second MC skill (GUESS), in a case where the agent also gives suggestions of guesses.

propositions, in addition to their utterances described above and in Figure 5.6: a list of uncertainties for the IDENTIFY agent, of guesses for the GUESS agent and of questions for the SEEK agent. Children were free to either use the agents' propositions or to enter their own answers. During the 6 remaining tasks, children did not have this help automatically and had to explicitly ask for it if they couldn't complete the task on their own.

5.2.3 *Technical implementation*

The behaviors of the four agents in terms of selection of the appropriate prompts as well as proposing a list of options (if requested) are entirely predefined and hand-scripted. Indeed, the agents are connected to a database containing the text resources and, for each one, a list of corresponding prompts relating to the four metacognitive skills, i.e., the utterances for the conversational agents for each skill. These utterances consist of sentences to remind the definition of the skill, its importance and how to use it in a think-aloud approach. Each text is also associated with a list of 3 propositions for each skill. All of these resources have been hand-generated by the research team and validated by two teachers for their pedagogical relevance.

During the interaction, the agent's automaton composes the dialogue utterances in order to include the appropriate cuing strategies. As we did for the "KidsAsk" platform, we changed the utterances between the texts to avoid repetition in the different agents' dialogue: a replica is not executed if it has been delivered during the previous text [112].

This implementation has no natural language processing or generative artificial intelligence methods to manipulate the behaviors for the different agents. This means that participants did not receive any specific qualitative feedback about their input. The agents only acknowledged when participants validate an entry (i.e. complete a step in the learning cycle). All data entered by participants was saved in a local database and was only evaluated post-experimentation, during the data analysis phase.

We choose this manual method to govern the agents' behaviors in order to be able to study the efficiency of the pedagogical approach itself. Once validated, we aim to automate it using NLP methods in order to upscale the method and facilitate its implementation! for various teaching activities.

5.3 EXPERIMENTAL PROCEDURE TO EVALUATE "KIDSREFLECT"

5.3.1 *Experimental conditions*

This study aims to design and evaluate the impact of a pedagogical training targeting perceptions of curiosity and specific metacognitive skills on children's metacognition, perception of curiosity, and divergent QA behaviors both during trainings and offline, spontaneous settings.

We choose to have three experimental groups: one that has the entire "KidsReflect" training (i.e. declarative via the videos + procedural via the web platform) as described above, one that only has the first part (i.e. only the declarative part) and one that has no "KidsReflect" training. To evaluate the difference between the three groups in terms of question-asking behaviors, we add a part where children work on a divergent question-asking task, using the WS2 of the "KidsAsk" platform we described in Chapter 3. During this task, children use a web-based platform to generate 18 questions about three different texts. For each question, a hand-crafted incentive conversational agent appears to give them two cues to help formulate their question: a semantic cue (a short sentence that represents a specific knowledge gap about the text) and a linguistic cue (a questioning word to help transform this KG into a question). See Figure 3.3 for an overview of the study timeline and the three experimental groups.

Concretely, our groups were like the following :

- Group ATEL (declarative + procedural) : had a 4-session declarative training that contained the 4 videos presenting curiosity and metacognition described above. They also had 2 sessions of procedural training using the web-based platform to practice the four MC skills. Finally, they also had two other sessions for the divergent question-asking practice, using "KidsAsk" with the hand-crafted incentive agent. Participants in this group had a total of 8 training sessions.

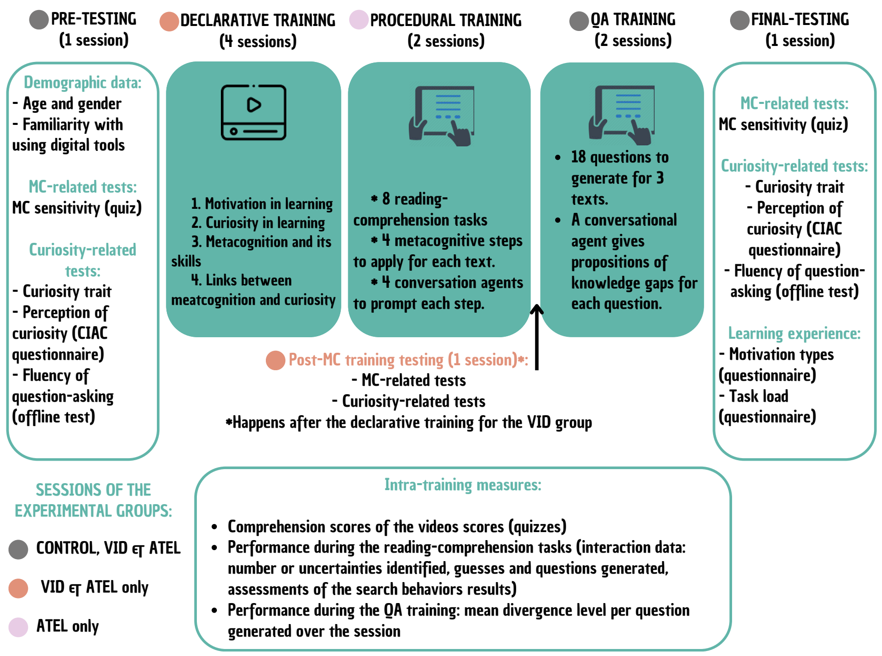


Figure 5.8: Summary of the "KidsReflect" training and the associated measures

- Group VID (declarative only) : had the exact same 4 sessions of declarative training related to curiosity and metacognition, as the ATEL group. These sessions were then followed by two others for the divergent question-asking practice, using "KidsAsk" with the hand-generated incentive agent. Participants in this group had a total of 6 training sessions.
- Group CONTROL (QA only): had two sessions for the divergent question-asking practice, using "KidsAsk" with the hand-generated incentive agent. This part is identical for the three groups. Participants in this group had a total of 2 training sessions.

The motivation behind this choice of groups is two-fold: 1) evaluate the impact of the declarative part alone on metacognitive- and curiosity-related measures and, in a second step, investigate the need to have the additional procedural training. Indeed, we want to propose a simple intervention that can be easily implemented in the classrooms by teachers, without the need for the research team to intervene. Our first part of the training answers this need as it consists of simple videos that can be easily discussed with students. And 2) the condition with only the divergent question-asking practice will serve us as a "control" condition to investigate the need for an introduction to curiosity and metacognition, prior to practicing and training divergent QA skills in order to enhance the efficiency of the latter. In other words, we investigate the importance to add a training such as "KidsReflect" in enhancing the efficiency of divergent QA trainings such as "KidsAsk".

5.3.2 *Procedure*

The "KidsReflect" intervention was composed of 6 45-minutes sessions for the Group ATEL (declarative + procedural) and 4 45-minutes sessions for the Group VID (declarative only). Both groups then had two 45-minutes sessions for the divergent QA training using "KidsAsk". The CONTROL group only had these two latter sessions. As illustrated in the timeline Figure 5.8, participants of all groups also had additional sessions that were used for the pre- and post-intervention measures and tests. Additionally, the ATEL and VID groups also had an immediate test session after they finished their "KidsRfelect" training, using the exact same tests.

Overall the ATEL group had 11 sessions (including 3 for the tests), the VID group had 9 sessions (including 3 for the tests) and the CONTROL group had 4 sessions (including 2 for the tests). See details about the measures and procedures to collect them in Section 5.4.

5.3.2.1 *The declarative knowledge sessions (for the ATEL and VID groups):*

During each of these sessions (4 in total), we present a 4-minute animated video explaining a concept related to curiosity and/or metacognition that we created within the team, as described in the section 5.2. During each session, the video is played a first time for the whole classroom and is then discussed with the research team: answer the students' questions, give further explanations for some concepts, etc. The video is then played for the second time and is followed by a second Q&A session. Finally, the session finishes with a 6-item quiz related to the video. These quizzes will serve us to evaluate the level of understanding, and thus of accessibility, of the concepts we presented to children.

An overview of the content of these videos and links to play them (in French) are available in Figure 5.2. The story board with the detailed scenario of one of the videos is available in Figure 5.3; the story boards for the rest of the videos are available in Appendix C.1 and the content of the quizzes used to evaluate the videos comprehension are available in Appendix C.2.

5.3.2.2 *The procedural knowledge sessions (for the ATEL group only):*

During these two sessions, the research team presents the platform of the training, described above in Section 5.2.2.2 and performs a demonstration to explain the different features and steps to follow. To make sure participants understood the task, the demo was followed by an example text to show the interaction with the four agents representing the MC skills. After this, children received each an individual tablet with the platform running and proceeded to do the task for a total of 8 texts over the two sessions.

For each text, once participants finish reading it they click on the "I've finished reading' button and the first referee appears (representing the IDENTIFY skill) and guides them into identifying a knowledge gap. After validating this step (through submitting an uncertainty), the detective agent appears to guide them into formulating a hypothesis about it (GUESS skill), and so on. As specified above (Section 5.2), the agents have predefined scripted behaviors and do not give any specific feedback regarding children's inputs apart from acknowledging the completion of the steps. See Figures 5.5 and 5.7 for an illustration of the interaction with the agents and a snap of the platform, and Figure 5.6 for a detailed example of the agents' utterances for one text.

During the first two cycles, and in order to help children get familiar with the task, the agents gave a list of three propositions along with the description of their roles (i.e a list of uncertainties for the IDENTIFY agent, of guesses for the GUESS agent and of questions for the SEEK agent); children were free to either use the agents' propositions or to enter their own input. During the 6 remaining tasks, children did not have this help automatically and had to explicitly ask for it if they were unable to solve the task themselves.

Two members of the research team were present all along the sessions to make sure to respond to any technical problem encountered by the participants.

5.3.2.3 *The divergent QA practice (for the three groups):*

During these two sessions, we started by introducing the "KidsAsk" platform and the task to participants. Participants only used the WS2 of KidsAsk, i.e. the workspace where they work on generating divergent questions for different texts, while receiving linguistic and semantic cues from a conversational agent to help them do so. In total, they had to work on 3 different texts to generate 18 questions.

Exactly as described above in Section 4.2.1, participants here interacted with a hand-crafted conversational agent that helped them generate questions by suggesting possible knowledge gaps about the text and giving a questioning word that they could use to transform this knowledge gap into a divergent question.

5.3.3 *Participants*

86 students aged between 9 and 11 were recruited from public primary schools in Bordeaux, France, with 44 girls and 42 boys. 33 participants were attributed to the ATEL group (declarative + procedural), 29 for the VID group (declarative only) and 24 for the CONTROL

Table 4: Profile measures for the participants

Measure	Group ATEL (M±std)	Group VID (M±std)	Group CONTROL (M±std)	p-value
Age	10.18 ± 0.63	9.63 ± 0.48	9.4 ± 0.3	0.08
Curiosity trait	27.93 ± 4.11	29.4 ± 4.79	27.2 ± 4.29	0.2
Initial divergent QA fluency	1.79 ± 1.19	1.91 ± 1.12	1.68±1.38	0.68
Initial CIAC score	33.42± 4.72	34.36 ± 5.93	43.45±7.6	0.3
Initial MC sensitivity score	0.7±0.09	0.7 ± 0.07	0.64±0.15	0.79

group (QA only). As it can be seen in the table below(Table 4), participants from the three groups were balanced with respect to the initial profile measures (details about the measures can be found in the following Section 5.4).

The study was approved by the institute's ethics committee (certificate n°2019-23) and only started after collecting all participants legal representatives' signed consents which contained the study goals, procedure and the data collected.

5.4 MEASURES

5.4.1 *Training measures*

In order to have a quantitative understanding about the accessibility of the training, i.e. to which extent participants were able to understand the learning steps and strategies proposed in the web platform and/or the content of the content, we investigate the two following measures:

- The comprehension score of the videos using a quiz about the video key content that participants had to complete after every video. Items of the quizzes for the four videos can be found in Appendix C.2.
- The performance during the web-based app for the MC training: the input of children was annotated by hand by the research

team to compute the percent of complete and correct learning cycles achieved. A learning cycle is considered correct and complete if: 1) the child entered the four steps. 2) for each step, the entry is directly related to the topic of the text. 3) the four steps are considered correct by the annotation grid. And 4) the four steps are related to each other semantically, i.e. they treat the same knowledge component. Details and examples of the manual annotation process for the four steps can be found in Appendix B.5.

5.4.2 *Outcome measures*

5.4.2.1 *Metacognitive sensitivity index*

A quantitative measure of metacognitive sensitivity, i.e. knowledge of one's own cognition, can be seen as the degree of association between one's accuracy in a task and their judgment of this performance; this measure is commonly known as "type 2" sensitivity. It is generally suggested that when individuals are endowed with metacognitive sensitivity, they will be more confident when they are correct. However, using simple correlation coefficients (e.g. Pearson's r) to quantify this may infer different biases [154].

Thus, and taking inspiration from the signal detection theory, we measure this variable with the type 2 ROC function (as suggested in [63]). This measure builds upon the widely used area under the receiver operating characteristic curve (AUROC) that uses participants' accuracy of confidence judgement in predicting performance. After AUROC is calculated, AUROC2 is then computed by adjusting AUROC for chance level. The closer this measure is to 1, the stronger a participant is in accurately evaluating their own knowledge. This is a key measure for us as we know that the accuracy in self-evaluation is crucial for introducing curiosity-driven learning cycles [147].

To collect this measure, we administer the same 12-item general knowledge quiz for all groups. For each item, they need to report how confident they feel that they have given the right answer using a 5-point Likert scale. In order to study the impact of the "KidsReflect" training , the "KidsAsk" training and both combined, on this index, we give the quiz several times: before the beginning of any intervention for all groups, after "KidsReflect" for the ATEL and VID groups, and after "KidsAsk" for all groups. We changed the quiz items across time so children do not see the same questions. Participants from all groups had the same questions at each time point.

Items of the quizzes used can be found in Appendix A.11.

5.4.2.2 *Perception of curiosity and QA*

In order to assess if and how children's attitude towards curiosity in general and asking questions in particular has evolved across time, we use Post's validated CIAC questionnaire [183].

The CIAC questionnaire contains sub-scales relating to: relation of social matters to curiosity, association of epistemic questions with curiosity, personal inclination, social relevance, negative opinion, fear of classmates negative judgement and self-efficiency. Items of the questionnaire used are available in Appendix A.2.

Indeed, this is an important measure for us as we think that one of the relevant brakes that can keep children from asking questions in the classroom is their negative perception of this behavior. Furthermore, seeing a significant change in this measure could mean that children had an important enough change in their metacognitive knowledge about curiosity and QA, that it affected their social perceptions of them.

5.4.2.3 *Divergent QA behavior during online practice sessions*

The performance during the divergent QA training: this measures consists of calculating the average divergence score per question generated during the training. A question's divergence score can range from 0 to 3. The specific procedure followed to calculate this score is described in Appendix B.3.

5.4.2.4 *Spontaneous divergent QA behavior*

This is an important measure for us as several research reports the tight links between spontaneous divergent QA and curiosity [4, 8, 69]. To measure this, we develop an offline test during which we present participants with a short text. We then give them 2 minutes to ask as many questions about it as they can/ want. Our aim being to compare this measure before and after the training sessions. This measure consists of calculating the mean divergence score of the correct questions children ended up generating. Similar to before, we assign the divergence score of the questions using the 4-point Likert scale described above and in Appendix B.3.

Children have this test before the beginning of any intervention for all groups, after "KidsReflect" for the ATEL and VID groups, and after "KidsAsk" for all groups. We give the same text for all participants, and only change it across time. The texts used can be found in Appendix A.10.

5.5 Results

During this section, we aim to study the efficiency of the "KidsReflect" training, in its two forms (declarative or declarative + procedural), on children's metacognition (through calculating their metacognitive sensitivity index) as well as on three main curiosity-related measures: 1) the perception of curiosity and QA, 2) the ability to benefit from a training targeting high-level and divergent QA skills and 3) the ability to transfer this training in new contexts to conduct spontaneous divergent QA behaviors.

In a second step, we also study the interaction between these different measures in order to investigate the factors suspicious of enhancing children's spontaneous curiosity-driven/ divergent QA behaviors. A summary of the measures and analysis we conduct is available in Figure 5.9.

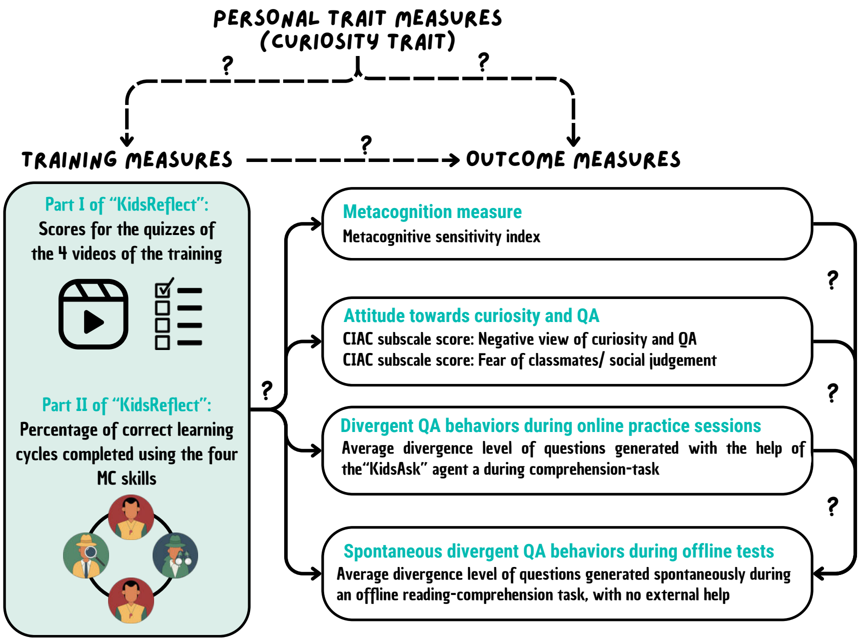


Figure 5.9: Summary of the measures and data analysis to be conducted

5.5.1 *Training measures*

Part I of the "KidsReflect" performance: understanding of the videos As explained above, the comprehension level of the videos was assessed by designing a 6-item quiz for each one and assessing participants' scores. Results revealed an high and similar comprehension of the videos content for the ATEL and VID groups (t=0.22, p-val=0.82): the verage comprehension scores for the 4 videos was of M=87.72%, SD=12.28 for the ATEL group and M=86.97%, SD=

Table 5: Children from both ATEL and VID groups had high comprehension scores for the four videos

Video	Group ATEL (M±std)	Group VID (M±std)	p-value
Video 1	82.06 ± 21.9	88.19 ± 18.5	0.24
Video 2	84.34 ± 15.86	81.95 ± 22.13	0.62
Video 3	90.3 ± 12.93	90.1 ± 17.18	0.96
Video 4	94.19 ± 11.28	87.64 ± 17.28	0.08

13.18 for the VID group. Same results were also found during the analysis per video (see Table 5 for details).

5.5.1.1 *Part II of the "KidsReflect" performance: completing the online tasks*

In evaluating the accessibility of the "KidsReflect" platform for the ATEL group, we investigate children's performance as a percentage of correct learning cycles using the four metacognitive skills appropriately. As explained in Section 5.4, for a cycle to be correct, the 4 steps should also be coherent, well-structured and targeting the same piece of information.

Our results show indeed good comprehension of the task as we register rather good scores and above average over the 8-texts training: M=77.39% and SD=15.8 of correct cycles.

5.5.2 *Outcome measures and impact of the training*

For the rest of the following section, we call an "Initial measure": a measure we take before any intervention, a "Mid measure": a measure we take after the "KidsReflect" sessions; whether in its declarative only or declarative + procedural form and, a "Final measure": a measure after the "KidsAsk" sessions, i.e. after all intervention sessions.

5.5.2.1 *Metacognitive sensitivity index*

As explained in Section 5.4, we calculate this index through the results of a quiz where children answer 12 questions and report their confidence levels in each answer. Children from the ATEL and VID groups complete this quiz three times (with different items every time): they have an initial, mid and Final measures. Children from the CONTROL group only had the Initial and Final measures.

In analyzing the difference in this score between the three conditions and within time, we first run a two-way ANOVA test. Results only show a tendency for an effect of the time on this score

($F(1,83)=3.83$, $p=0.054$). However, we see no interaction effect between the time and the condition on the scores: $F(2,83)= 0.029$, $p=0.74$. See Figure 5.10.

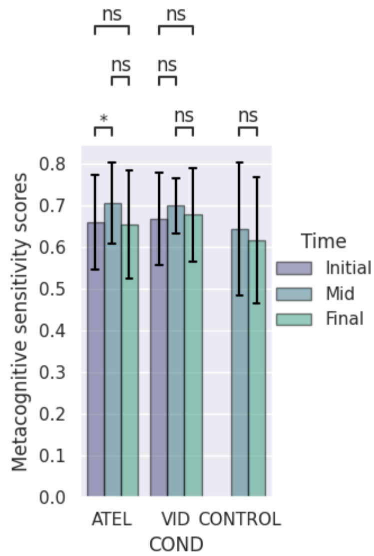


Figure 5.10: Only children from the ATEL group were able to significantly enhance their MC sensitivity after the KidsReflect" training. The effect is lost after "KidsAsk".

We then proceeded to perform pairwise comparisons with Bonferroni corrections to disentangle this difference seen over time. Results show that:

- For the ATEL group: a significant difference is seen between the initial and mid measures: p-adj=0.05. However, only a tendency for a difference is seen between the Initial and Final measure (p-adj=0.08) and no difference is seen between the mid and Final measures (p-adj=0.8).

- For the VID group: a tendency for a significant difference is seen between the initial and mid measures: p-adj=0.06. However, no difference is seen between the Initial and Final measures (p-adj= 0.3), nor between the mid and Final ones (p-adj=0.6).

- For the CONTROL group (only has the initial and Final measures): no difference is seen between the two measures (p-adj= 0.4).

Finally, in analyzing the impact of the training on this measure, we conduct 2 mixed ANCOVA analyses with the MC sensitivity score across time as an outcome variable for the ATEL and VID groups: one with the average videos comprehension score as dependent variable for both groups, and the other with the the percentage of learning cycles completed during Part II of "KidsReflect" for the ATEL group. Results show two significant interactions: F=26.56 and p<0.0001 for the videos score and F=6.88, p=0.01 for the platform score, thus suggesting the efficiency of the two parts of "KidsReflect" in enhancing metacognitive efficiency.

Taken together, these results suggest a tendency for an enhancement in children's metacognitive sensitivity, but only for those who had the "KidsReflect" training and with a stronger effect size for the ATEL group who had the full training, i.e. declarative and procedural metacognitive knowledge about curiosity. Indeed, these enhancements were correlated to the performance during the two parts of the training. However, our results also suggest that these effects do not last in time and are not maintained through the "KidsAsk" sessions, as we see no differences between the Initial and Final or the Mid and Final measures for any group.

5.5.2.2 *Perception of curiosity and QA*

In analyzing children's perception of curiosity, we use the CIAC questionnaire developed in [183] and explained in Section 5.4. Our goal is to investigate the difference in this measure before and after the different components of our intervention, between our three groups. To do this, we start by running a mixed ANOVA test with the CIAC score as a dependent variable, between groups and within time. Our results show indeed a significant interaction effect between these two latter on children's reported CIAC score: F(2,83)=27.44, p-val<0.0001). In doing pairwise investigations, we find indeed the same interaction effect between the ATEL and the CONTROL groups: F(1,55)=43.97, p<0.0001; as well as between the VID and the CONTROL groups: F(1,49)=35.92, p<0.0001. However, this interaction is not significant between the ATEL and VID groups (F(1,58)=0.14, p=0.86).

In a second step, we investigated the dimensions "negative view of question-asking" and "fear of negative judgement", 2 sub-scales in the CIAC questionnaire that can have an important effect on curious behaviors in the classroom [69].

- For the negative view of curiosity: we only see a significant effect of time (F(2,83)=14.13, p<0.0001). In doing pairwise comparisons with Bonferroni corrections, we only see a difference between the initial and the Mid measures (p-adj<0.0001). In investigating this difference deeper, we see that only the ATEL group had a significant drop between the Initial and Mid measures (p-adj=0.0002) as well as between the Initial and Final measures (p-adj=0.0003); no difference is seen between the Mid and Final measures (p-adj=0.8). See sub-figure (a) in Figure 5.11.

- For the fear of negative judgement: here again, we only see a significant effect of time (F(2,83)=18.95, p<0.0001). In doing pair-wise comparisons with Bonferroni corrections, we see a difference both between the initial and the Mid measures (p-adj=0.0002) and the Initial and Final measures (p-adj=0.001). In investigating this difference across the groups, we see that both children from the ATEL and VID groups were able to decrease their score of fear between the Initial and Mid measures (p-adj=0.02 for the ATEL group and 0.008 for VID), as well as between the Initial and Final measures (p-adj=0.008 for the ATEL group and 0.002 for VID). However, no difference is seen between the Mid and Final measures (p-adj=0.67 for the ATEL group and 0.62 for VID). No difference is seen between the Initial and Final measures for the CONTROL group (p-adj= 0.32). See sub-figure (b) in Figure 5.11.

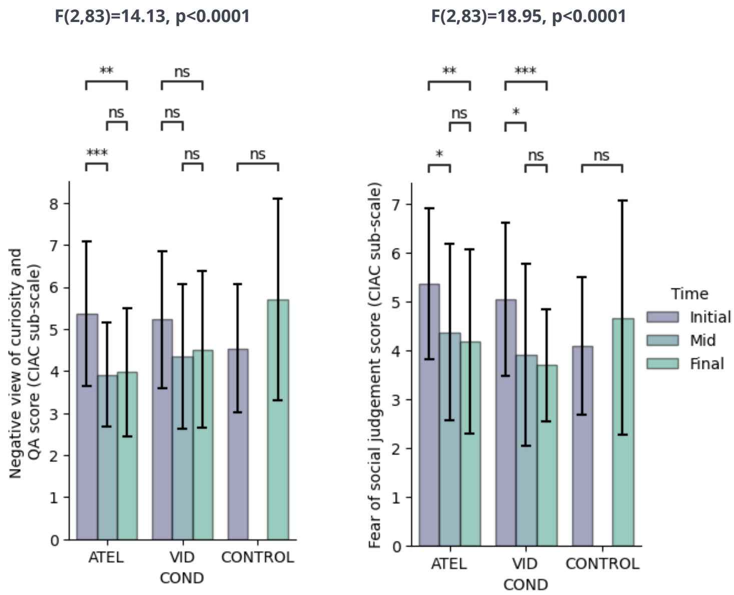


Figure 5.11: Only participants from the ATEL group were able to enhance their perception of curiosity and QA. Participants from the ATEL and VID groups were both able to reduce their fear of judgement when asking questions. Participants from the CONTROL group were not able to change these dimensions.

Finally, in analyzing the impact of the training on the change we see in the CIAC scores across time, we conduct 2 mixed ANCOVA analyses for the ATEL and VID groups: one with the average videos comprehension score as dependent variable for both groups, and the other with the the percentage of learning cycles completed during Part II using the "KidsReflect" platform for the ATEL group. Results show a significant interaction for the videos score: F=19.62, p<0.0001, but not for the platform score: F=0.06, p=0.8.

Taken together, our results suggest the efficiency of having simple video-based interventions such as Part I of "KidsReflect" to enhance children's perceptions of curiosity and especially decrease their fear of negative judgement when adopting such behaviors in the classroom. Indeed, we see a correlation between the enhancement in the CIAC score and children's good understanding and adoption of the instructional knowledge about curiosity given during part I of "KidsReflect". Having question-asking trainings alone (such as "KidsAsk") does not seem to help them change these perceptions.

5.5.2.3 *Divergent QA behavior during online practice sessions*

One of our hypotheses is that "KidsReflect" can help favour the efficiency of divergent question-asking trainings such as "KidsAsk", i.e., it will help children benefit better from this latter.

To investigate this hypothesis, we assess children's performance during a specific task in the "KidsAsk" platform we presented in Chapter 3: in this task, children had to generate 18 questions in total about 3 different texts. For each question, they had the help of a handcrafted agent that gave them propositions of knowledge gaps and question starters to push them towards transforming the KGs into divergent questions.

We assess their performance during "KidsAsk" using exactly the same methods we described in our earlier studies and in Section 5.4.2, i.e., we compute the average divergence score of the questions generated during the sessions. Details about the procedure of how to calculate this score are available in Appendix B.3.

We start by performing a one-way ANOVA test between the three groups; results show indeed a significant difference in the performances (F(2,83)=6.45, p-value=0.002). We then perform pairwise T tests with Bonferroni corrections to understand the differences between the groups. Results show no significant difference between the ATEL and the VID groups (t=-1.11, p-adj=0.3 and power=0.2). However, the CONTROL group had a significantly lower performance compared to the ATEL group (t=-3.53, p-adj=0.001 and power=0.95) and the VID one (t=-2.6, p-adj=0.03 and power=0.74). See sub-figure (a) in Figure 5.12.

Finally, we also analyze the impact of the "KidsReflect" training on this measure for the ATEL and VID groups. We thus run two correlation analysis : one with the average videos comprehension score as dependent variable for both groups, and the other with the the percentage of learning cycles completed during Part II using the "KidsReflect" platform for the ATEL group. Results show a significant correlation for the videos scores: F=19.93, p<0.0001 and for the platform

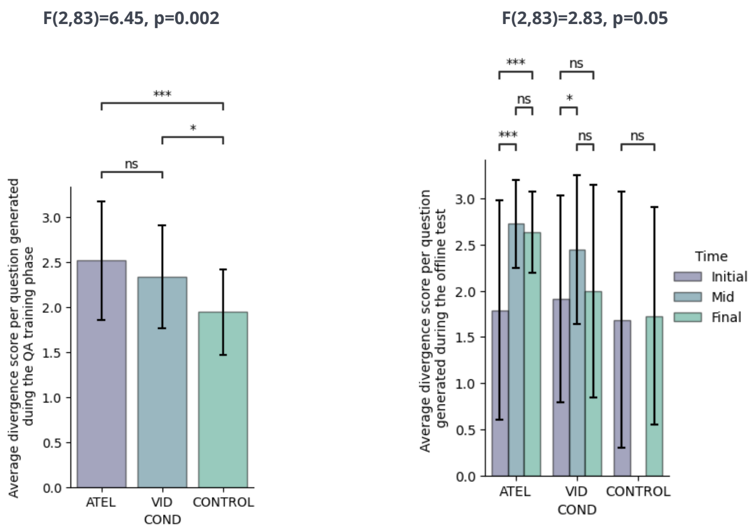


Figure 5.12: (a)Participants from the CONTROL group benefited the least from the divergent QA practice sessions with "KidsAsk". The ATEL and VID groups had similar behaviors. (b) Participants from the ATEL group had significantly more progress in their spontaneous divergent QA behaviors than the two other groups.

score: F=33.31, p<0.0001, thus suggesting the efficiency of "KidsReflect" in its two parts in helping children benefit from "KidsAsk".

With these results, we can suggest that children who had the "KidsReflect" training, whether in its declarative only or declarative + procedural form, were able to benefit better from the divergent QA practice sessions. The performance seems to be indeed directly correlated to their understanding of "KidsReflect".

5.5.2.4 *Spontaneous divergent QA behavior*

In analyzing children's spontaneous divergent QA fluency, we run an offline test where children have to generate questions with no external help, during a reading-comprehension task as described in Section 5.4.2. We then calculate the average divergence level per question generated during this test, following the same procedure we use before and report in Appendix B.3.

Once calculated, we run a two-way ANOVA test to investigate whether we see a change in this measure over time, between the three groups. The results show indeed a significant interaction effect between time and the experimental group on children's score in the test: F(2,83)=2.83, p-value=0.05 (see sub-figure (b) in Figure 5.12). In doing pairwise comparisons with the Bonferroni corrections, we only see a significant enhancement in this score for the ATEL group: between the Initial and Mid measures (p-adj<0.0001) as well as between the Initial and Final measures (p-adj<0.0001) but not between the Mid and Final measures (p-adj=0.65). However, we see no differences for the VID group between the Initial and Mid measures (p-adj=0.17) nor between the Initial and Final measures (p-adj=0.76) or between the Mid and Final measures (p-adj=0.32). Finally, no difference is seen for the CONTROL group between the Initial and Final measures: p-adj=0.9.

Given the important variance in this measure (as seen in the Figure 5.12), we also analyze the standard deviations between the three measures, across the three conditions, using F-tests. Our results showed indeed a significant decrease in the standard deviations for the ATEL group: p<00001 between the Initial and Mid measures, p-val<0.0001 between the Initial and Final ones. No difference is seen between the Mid and Final measures (p-val=0.67). However, results ware different for the VID group: a significant decrease is seen between the Initial and Mid measures (p-val=0.04), no difference is seen between the Initial and Final measures (p-val= 0.55) and we see an increase between the Mid and Final tests (p-val=0.03). Finally, no significant differences between the standard deviations are seen for the CONTROL group (p-val=0.22 between the Initial and Final measures).

On a final note, and in analyzing the impact of the "KidsReflect" training on this measure for the ATEL and VID groups, we run two correlation analysis : one with the average videos comprehension score as dependent variable for both groups, and one for the ATEL group with the the percentage of learning cycles completed during Part II of "KidsReflect" platform. Results show a significant correlation with the videos score (F=4.68, p=0.03) and a tendency for a correlation with the platform score (F=2.95, p=0.07).

Taken together, these results suggest indeed the importance of trainings such as "KidsReflect" that combine giving theoretical and experience-based knowledge about curiosity and metacognition: it seems to consolidate children's ability to ask more spontaneous divergent questions, with no external help, and when they're in a new learning contexts. In other words, this metacognitive training with 2 components seems to have a stronger transfer effect of the divergent QA skills and to help reduce the gaps between children in this ability.

5.5.3 *Interactions between the training and outcome measures*

In a final step of our analysis, we aim to make sense of the changes seen in children's metacognition- and curiosity-related measures by studying the interactions between them. This is a crucial step for our study as it will help give research guidelines with respect to the dimensions needed in order to implement efficient training for curiosity-driven QA behaviors during learning.

To do this, we run different step-by-step linear mixed regression models. For each model, the dependent variable represents the Final measure for one of our outcome variables. The list of candidate predictors used to run the models comprises interpersonal measures (i.e. curiosity trait), the training measures, and the outcome measures we had during the Initial, Mid and Final tests. See Figure 5.13 for a summary of the measures given as predictor candidates for the mixed linear models we run.

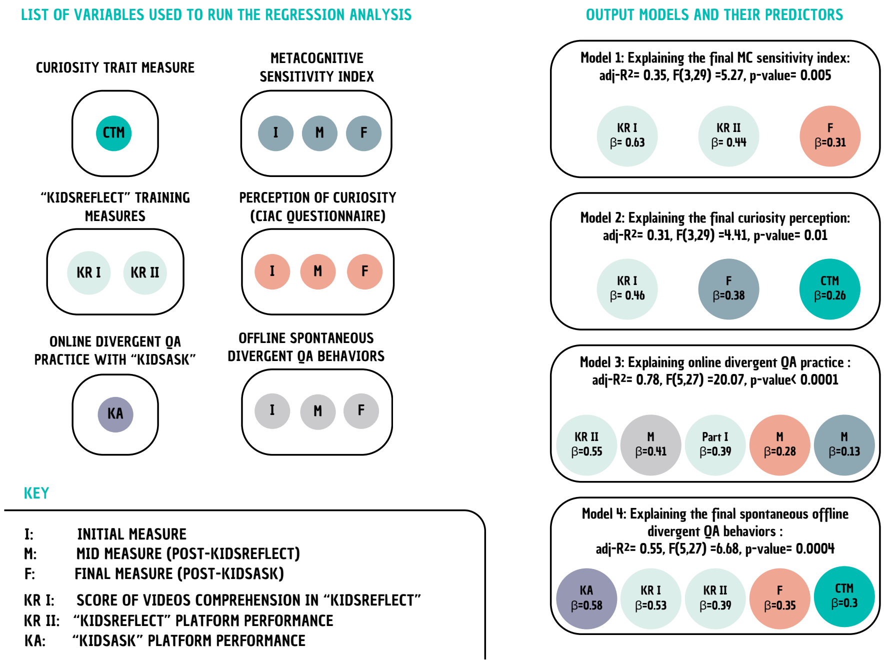


Figure 5.13: Summary of the regression analysis run to understand the final scores of the outcome measures along with a full list of input variables used for the investigation.

5.5.3.1 *Explaining the final metacognitive sensitivity index*

We start by investigating the factors that influenced children's metacognitive sensitivity measure after the full intervention. We run a step-by-step regression analysis with the Final measure of this score as a dependant variable.

We found an overall statistically significant regression: adj-R^2= 0.35, $F(3,29)$ =5.27, p-value= 0.005. The factors found to predict this score are: the average score of the videos comprehension during the part I of the "KidsReflect" training (β=0.64, p=0.002), the percentage of correct learning cycles completed during the part II of the "KidsReflect" training (β=0.44, p=0.02) and the Final measure of the perception of curiosity score (β=0.31, p=0.044). A list of the potential candidates passed for this model as well as a summary of the results can be seen in Figure 5.13, "Model1".

These results, along with the significant correlations found with the training measures in Section 5.5.2.1, suggest indeed the importance of the two parts of the "KidsReflect" training on children's metacognitive sensitivity. Having an enhanced perception of QA seems to impact this measure as well.

5.5.3.2 *Explaining the change in the perception of curiosity and QA*

In trying to understand the factors involved in the enhancement we saw in children's perception of curiosity and QA, we also run a step-by-step regression analysis with the Final measure of children's CIAC score as a dependent variable.

We find a statistically significant regression : adj-R^2= 0.31, F(3,29) =4.41, p-value=0.01. The model indicates that the factors found to predict the final measure of children's perception of curiosity are: the average score of the videos comprehension during the part I of the "KidsReflect" training (β=0.46, p=0.01), children's Final metacognitive sensitivity index, i.e. after "KidsAsk" (β=0.38, p=0.012) as well as their curiosity trait (β=0.26, p=0.1). Results of the model as well as the list of the all input variables can be seen in Figure 5.13, "Model2".

These results, along with the significant correlation found with the videos scores in Section 5.5.2.2, suggest the importance of our part I of the training, children's metacognitive sensitivity and their curiosity trait (as reported by the parents) on their perception of curiosity and QA. This suggests that these perceptions, even though can partly depend on personal traits, can be changed using specific and simple interventions such as videos, and that focus on giving theoretical knowledge about curiosity and metacognition.

5.5.3.3 *Explaining the divergent QA behaviors with "KidsAsk"*

To investigate the factors impacting children's performance during the divergent QA training with "KidsAsk", we run the regression analysis with the performance score during these sessions as a dependant variable. We find an overall significant regression (adj-R^2=0.78, F(5,27)=20.07, p-value<0.0001). The factors found to predict this score were: the percentage of correct learning cycles completed during the part II of the "KidsReflect" training (β=0.55, p<0.0001), the Mid score of the spontaneous divergent QA fluency (β=0.41, p<0.001), the average score of the videos comprehension during the part I of the "KidsReflect" training (β=0.38, p=0.004), the Mid score of the perception of curiosity and QA (β=0.28, p=0.005) and the Mid score of the MC sensitivity index (β=0.13, p=0.1). Results of the model as well as the list of the all input variables can be seen in Figure 5.13, "Model3".

5.5.3.4 *Explaining the change in the spontaneous divergent QA behaviors*

We investigated the progress in the offline divergence QA fluency overtime. The aim being to understand what can lead children to adopt these behaviors spontaneously, in different learning contexts and when they have no external help or incentives to do so (contrary to the settings in "KidsAsk").

Here again, we run a a step-by-step regression analysis with the score in the final offline divergent QA test as a dependant variable. We find a significant regression (adj-R^2=0.55, $F(5,28)$=6.68, p-value= 0.0004). The factors found to predict this score are: the performance during the "KidsAsk" practice sessions (β=0.58, p=0.006), the average score of the videos comprehension during the part I of the "KidsReflect" training (β=0.53, p=0.002), the percentage of correct learning cycles completed during the part II of the "KidsReflect" training (β=0.39, p=0.05), the final score of curiosity perceptions (β=0.35, p=0.014) and children's curiosity trait (β=0.3, p=0.05). Results of the model as well as the list of the all input variables can be seen in Figure 5.13, "Model4".

Since we see no direct link in these results between the spontaneous divergent QA behavior and metacognition, we run an additional mixed ANCOVA test with the Final measure of metacognitive sensitivity as an independent variable in order to investigate its impact on the divergent QA score within time. Results show indeed a significant interaction: F=4.71, p-val=0.03.

These results, along with the significant correlation found with the videos scores found in Section 5.5.2, suggest that children who had "KidsReflect" training in either one of its forms tended to have better progress in their ability to conduct spontaneous divergent QA behaviors, compared to those who only had the "KidsAsk" sessions. This progress tended to be stronger for those who had the full training (i.e. the ATEL group). Furthermore, the final measure of this skill seems to be predicted by the ability to benefit from the different interventions (i.e. "KidsReflect" with its two parts and "KidsAsk") as well as metacognitive sensitivity, perception of curiosity and curiosity trait.

A summary of these findings can be seen in Figure 5.14.

5.6 discussion

In this study, we aimed to explore the impact of teaching children about curiosity, its significance as an effective and enjoyable learning strategy, and how to control it through the practice of specific metacognitive skills. To achieve this, we designed a pedagogical intervention we call "KidsReflect." The training was developed in two versions: one that only provides conceptual/ theoretical knowledge

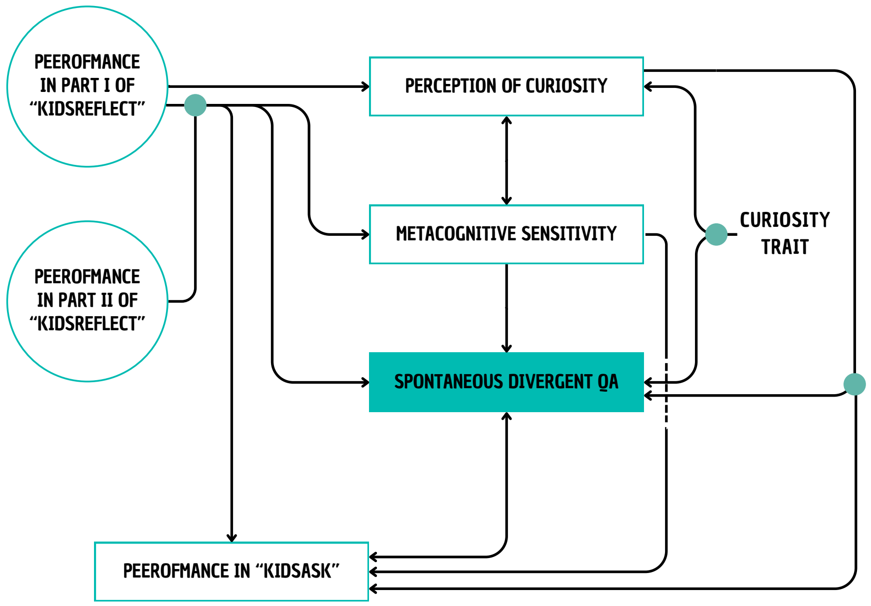


Figure 5.14: Summary of the results showing the different relationships between the training and outcome measures.

about curiosity and related metacognitive skills through animated videos created by our team. The second version added also a procedural training to apply these skills during a reading comprehension task. To do this, we designed a web platform that helps children practice these skills with specific cues provided by conversational agents.

We evaluated the effectiveness of the two versions of "KidsReflect" by examining their impact on children's metacognition (by measuring their metacognitive sensitivity index), and three main curiosity-related measures: 1) attitudes towards curiosity and QA behaviors in the classroom, 2) the ability to benefit from training aimed at enhancing high-level and divergent QA skills, and 3) the ability to transfer this training to new contexts to engage in spontaneous curiosity-driven QA behaviors, without any external help or incentives.

Our results showed that children who received the full "KidsReflect" training (i.e. had both the declarative and procedural parts of the training) generated significantly more divergent questions during an offline test, where no external help or incentives are offered. This suggests that, compared to those who has the declarative "KidsReflect" only or no "KidsReflect", these participants were the most adept at transferring the skills learned during the training to new contexts and at using them independently. Additionally, we find that the ability to autonomously ask divergent questions was correlated to the performance during the "KidsReflect" and "KidsAsk" trainings, as well as to the MC sensitivity, perception of curiosity, and curiosity trait.

5.6.1 *Summary of findings*

on the importance of linguistic qa trainings First, we that the performance during the "KidsAsk" training seems to be a predictor for the final spontaneous divergent QA behaviors. This result comes to reinforce previous findings showing the importance of interventions that reinforce children's semantic and linguistic QA skills—necessary to ask high-level curious questions [3, 6, 8]. It also supports several ideas that suggest linguistic QA skills as main barriers that keep children from asking questions [43, 69, 79].

These results are also in line with previous QA models such as in [169] suggesting the importance of language for QA behaviors. This importance is also highlighted in the model proposed by Ronfard et al. [192] where one essential step to ask questions if 'formulation', i.e. the ability to phrase a question in order for it to be understood.

Furthermore, we also find that in order for students to be able to benefit from this training, a good understanding of the two parts of "KidsReflect" seems important. This suggests indeed the role of reinforcing knowledge about curiosity and metacognition to favour the efficiency of divergent QA trainings.

on the importance of metacognition Second, we also see that metacognitive sensitivity was a predictor for children's final divergent QA behaviors.

This relationship aligns with he idea that curiosity heavily depends on metacognitive judgements, i.e. how much individuals know about their own knowledge [77, 142]. This is also in line with the knowledge gap theory where the level of information-seeking behaviors is linked to the individual's ability to detect a discrepancy in their knowledge [147]. Similarly, Ronfard et al. also suggest in their QA model that the first step in the process of asking a question is *initiation*, i.e. the "realization that information is lacking and needed" [192].

This may suggest that the more accurately participants judged their own knowledge levels (i.e., the more they were metacognitively efficient), the better they were at recognizing knowledge gaps they wanted to explore, and, thus, the more they tended to ask divergent questions.

Furthermore, given the direct links observed between the metacognitive sensitivity index and the children's performance and understanding of both components of "KidsReflect", we can infer that our training methods we rather efficient in providing children with the relevant metacognitive skills needed to facilitate their ability to ask divergent questions.

However, it is important to note that we only saw a slight significant increase in metacognitive sensitivity for the ATEL group and that this effect disappears after "KidsAsk". Several reasons can be advanced for this observation, mainly the short length of our "KidsReflect" training. Indeed, changing such a complex dimension may require longer-lasting and more domain-specific interventions [202]. It also possible that this is due to the lack of feedback we gave children about their performance during the training. Indeed, the task we proposed was relatively new to students (identifying uncertainties, generating guesses, etc) and different from the tasks they normally are asked to complete in schools. It is thus possible that not having any information about their performance may have prevented them from fully benefiting from the training. Finally, it is also important to note that we only assessed metacognition based on the sensitivity index, which may not be a sufficient measure as it only contains information about children's knowledge of their cognition and not their ability to regulate it —an equally important dimension to lead curiosity-driven learning processes.

ON THE IMPORTANCE OF CURIOSITY PERCEPTIONS In another important observation, we see that children's final spontaneous divergent QA behavior was related to their perception of curiosity. This interesting result gives empirical evidence to several assumptions that one major brake keeping children from asking questions in the classroom is their negative idea of this behavior (e.g. association with stubbornness, stupidity, etc) and their fear of negative judgement when asking questions [42, 69, 109, 183].

Participants from both the ATEL and VID groups were able to significantly increase their overall perception of curiosity after the "KidsReflect" training, unlike those from the CONTROL group. We also see that participants from the ATEL group had the most decrease in their negative view of QA behaviors (a sub-scale in the CIAC questionnaire).

Interestingly, these changes were also found correlated to the understanding of the videos presented during the first part of the "KidsReflect". This result suggests indeed to the importance for teachers to explain the nature and importance of curiosity and QA behaviors during learning for their in order to correct some misconceptions that students may have.

ON THE IMPORTANCE OF INDIVIDUAL DIFFERENCES Finally, we also saw that children's final spontaneous divergent QA performance was partly dependent on their curiosity trait (as reported by their parents). This result reproduces findings from previous studies [8] and reinforces the idea to consider divergent question-asking behaviors as an indicator for epistemic curiosity. This is an important idea as it suggests divergent QA performance as a substitute behavioral measure for curiosity rather than the self-report ones that are usually used and that have been long been questioned due to the different biases they can introduce.

As a conclusion, our results point out that having a pedagogical training that provides both theoretical and experience-based metacognitive knowledge of curiosity was the the most effective method for improving our targeted curiosity behavior, i.e. spontaneous divergent QA. This is inline with metacognition models such as in [61] suggesting that metacognition has two essential components: conceptual and experience-based MC.

5.7 LIMITATIONS AND FUTURE DIRECTIONS

Despite these promising results suggesting the need for declarative and procedural metacognitive knowledge to facilitate the training and enhancement of children's divergent QA behaviors, our methods remain rather limited on different dimensions.

First, the methods used to implement our intervention and the immediate testing approach we took are important factors that could have played a role in the positive results we saw. Indeed, since we had relatively short and novel, attractive interventions (video-based, tablet-based activities which remain new in the classroom), this may have helped engage children more in the activities, resulting in the positive results we saw. Furthermore, since we took all of our measures immediately after the intervention (one day after "KidsReflect" or "KidsAsk"), this may also have helped us see our positive results. It is therefore interesting to think about future implementations that could: 1) be integrated in the formal classroom settings and adapted by teachers during their standard pedagogical activities with students. And 2) to have an experimental design where we can measure the enhancement in curiosity-driven behaviors in a longer-term, e.g. over a school year. To do this, we are starting a collaboration with volunteer teachers in order to adapt the approach we presented in this study to the formal activities they normally carry on with their students. The goal is for them to be able to use this approach independently and be able to use it alone and adapt it to their teaching methods and pedagogical goals.

Second, we identified a technical limitation in our method regarding the lack of feedback on users' performance in both online trainings: the second part of "KidsReflect" and "KidsAsk". Indeed, the two platforms do not contain any natural language processing modules to assess the quality of children's input and give real-time feedback. Feedback is a crucial factor to maintain children's activeness and awareness of their learning process. It can become even more crucial and important for the novel tasks such as the ones we propose: children are not used to being asked to identify uncertainties, generate guesses or questions, or evaluate new pieces of information. Therefore, in future implementations, we also aim to add this dimension to our platform in order to give information about specific dimensions of children's answers (divergence level of a question, usefulness of an uncertainty given a specific activity, etc). One potential idea could be to use Large Language Models (LLMs) to generate such real-time feedback as previous studies have shown their efficiency in similar tasks when we use specific prompting methods [229, 231].

As mentioned above, we also have limitations in terms of assessing metacognition as we only capture the sensitivity index. To tackle this, we aim to combine this with other measures that include metacognitive learning strategies, self-regulation, etc using for instance the junior metacognitive awareness inventory [214] or the motivated strategies for learning (MSLQ) [181]. We also think that it could be interesting to analyze children's performance in part II of "KidsReflect" by accounting for the success in each of the four metacognitive skills. This could help us evaluate how each of these steps contributes to the divergent QA process.

On a final important note, we highlight the fact that, during this study, we do not have domain-knowledge learning progress measures. This is due to experimental design constraints (need for extra sessions and time constraints, etc). This is an important limitation in our study, considering that our ultimate goal is to evaluate the impact of training curiosity-driven and MC-regulated learning strategies as a means to enhance learning. This is indeed another motive for us to implement longer-lasting and more ecological implementations of "KidsAsk" and "KidsReflect", as they could allow us to investigate such a dimension.

5.8 conclusion

The current work aimed to understand new factors that could facilitate the training and enhancing of children's spontaneous curiosity-driven question-asking behaviors. More particularly, we focus on the role of metacogniton and perception of curiosity.

Taken together, the results of this study highlight the multidimensional factors that could influence this skill and suggests the efficiency of the methods we propose to enhance it: personal traits (i.e. curiosity trait), the perception of curiosity, metacognitive sensitivity and linguistic QA skills.

These results motivate further research to explore the complex interactions between curiosity and metacognition during learning and weigh in the importance of each MC skill in this process. It also motivates the integration of simple metacognition and curiosity knowledge in the classroom, to promote healthier perceptions of curiosity and QA.

Part IV

DISCUSSION

6

DISCUSSION

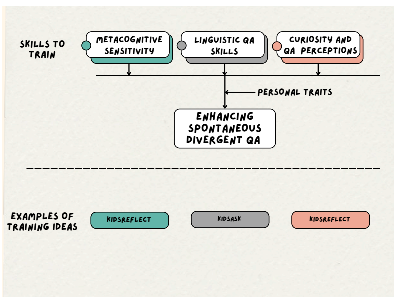
Collaborators: Xingdi Yuan, Q Vera Liao, Microsoft Research, Montreal. Celeste Kidd, University of Berkeley, California.

Aims: This chapter puts together the results obtained with the work of this thesis, and discusses their implications for future research around promoting curiosity and metacognition, especially in the age of Generative Artificial Intelligence.

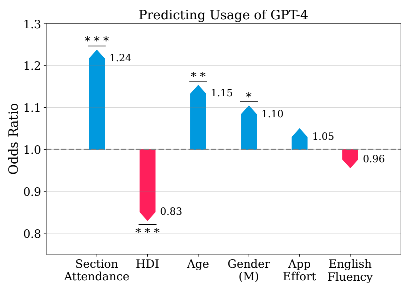
Content: The chapter begins with a summary and discussion of the main findings of this thesis. It highlights three key dimensions we find essential to facilitate the promotion and training of divergent question-asking behaviors: linguistic QA skills, metacognitive sensitivity and perceptions of curiosity. The chapter also addresses the primary limitations of our studies— particularly concerning the design and technical implementation of the training methods— and presents the perspectives and future directions of using these methods both in classrooms and in the EdTech industry. Finally, it asks an open question regarding the relevance of our work in the age of GAI and presents a new study we designed to investigate this.

Contents

The studies presented in the perspectives of this chapter are published in IUI'23: ***Ziang Xiao et al. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding." In:*** **Companion Proceedings of the 28th International Conference on Intelligent User Interfaces.** ***2023, pp. 75–78*** and in GAIED, NeurIPS'23 workshop: **Rania Abdelghani et al. "Generative AI in the Classroom: Can Students Remain Active Learners?" In:** ***arXiv preprint*** **(2023)**



(a) Functional ingredients found predictive of divergent QA



(b) Predictions of GPT-4 use in education

6.1 SUMMARY

The general aim of this thesis is to propose new educational technologies that can support and enhance students' curiosity through the practice of active learning strategies such as question-asking and self-evaluative and monitoring skills (i.e. metacognitive skills). Particularly, we aim to present three main contributions: 1) validate the efficiency of specific QA-supporting instructional design principles and their impact on learning and curiosity. 2) validate the usefulness of GAI to facilitate the implementation of curiosity-driven QA e-learning environments, and 3) validate the efficiency of specific training methods to enhance children's declarative and experience-based knowledge about curiosity and its related metacognitive skills.

In developing these instructional design principles, we first started by exploring curiosity definitions and models, triggering mechanisms, etc. Indeed, in Chapter 2, we start by reviewing the various definitions of epistemic curiosity and end-up introducing it as a specific form of intrinsically-motivated information-seeking behavior, mainly motivated by an expected informational gain. We then provide an in-depth insight into the mechanisms of epistemic curiosity, the frameworks proposed to explain the internal and external factors that could stimulate it, its forms of expression— mainly divergent question-asking— and its role in cognitive development.

We also identify a recurrent factor that seems to be tightly related to curiosity, which is metacognition i.e., the individual's knowledge about their own cognition. In the second section of our theoretical chapter, we thus introduce metacognition, its different components, forms and developmental trajectory. We also provide a discussion about the deep connection between specific metacognitive skills and curiosity-driven behaviors. Namely, we talk about explicit self-reflection, evaluation of one' own knowledge, identification of missing information, self-regulation of learning strategies to pursue an identified knowledge gap (KG), etc.

Finally, we also review curiosity-driven strategies in formal educational settings: their current use by the pedagogical teams in schools, the practical guidelines available for teachers today to promote curiosity in the classroom and the impact of this latter on students' learning experiences and outcomes. We also talk about the digital transformation in schools and the role of new technologies to help support teachers and students during the learning process. More specifically, we discuss the role of novel technologies, especially with natural language processing, in supporting students' metacognition and curiosity-driven learning strategies.

The empirical contributions of this thesis are then reported in Part iii. In this part, we present three studies. 1) **To investigate the efficiency of a training for divergent question-asking skills:** the first study reports empirical evidence for the efficiency of the instructional design principles we implemented in "KidsAsk". A platform aimed to train children's divergent QA skills by explicitly showing them their potential KGs and providing linguistic help to formulate well-structured questions directed towards the compensation of these KGs. To do this, "KidsAsk" offered interactions with a conversational agent that gives two specific cues to facilitate these two dimensions during a reading-comprehension task: a short sentence to propose knowledge gaps and a high-level questioning words to facilitate the question construction. The behaviors of the agent were completely generated manually beforehand, by the research and pedagogical teams. Additionally, "KidsAsk" also offered a space to investigate the impact of this training on children's spontaneous explorations and subsequent learning progress (see Chapter 3).

Our results showed the efficiency of this training in helping increase students' divergent QA behaviors, both during the training itself and during a different learning activity we proposed afterwards where there is no external help to formulate the questions. During this activity, we also find that participants who asked more divergent questions ended up having longer-lasting autonomous exploratory behaviors and stronger domain-knowledge learning progress. These findings suggest indeed the efficiency of our QA-training strategy and encourage its use in formal learning settings.

2) **To investigate the feasibility of using LLMs to generate the content for a divergent QA training:** in our second empirical contribution, we study the feasibility of using an LLM (GPT-3) to automate the implementation of a divergent question-asking training such as "KidsAsk" (see Chapter 4). In this study, we use specific prompting approaches to generate semantic and linguistic cues with GPT-3 in order to help children think of and formulate divergent questions. These cues are then proposed by our conversational agent in "KidsAsk" in order to practice generating divergent questions during reading-comprehension tasks.

Since we saw that the cues offered in Study 1 were rather efficient, we design our prompting approach with GPT-3 in order to generate cues that have the same structure. Additionally, we also use GPT-3 to explore generating other a new form of cues. More specifically, we generate cues with an open structure: a list of relevant keywords that could be linked together in different ways and lead to think of different divergent questions. Unlike the approach in Study 1, this one leaves children with more choice over the questions they can think of using the help of the CA.

The results of this study show the efficiency of using LLMs such as GPT-3 to generate the pedagogical content for divergent QA trainings. Indeed, we see no difference in children's behaviors between agents that proposed hand-generated vs. GPT-3-generated 'closed' cues. However, we see a significantly better performance for those who had the more 'open' cues and, unlike for the 'closed' ones, we see a correlation between the QA performance and the curiosity trait. These results suggest the relevance of this new strategy to better support curiosity-driven question-asking.

And finally, 3) **To investigate the efficiency of a adding a training for metacognition and curiosity perceptions:** in our third empirical contribution, we try to leverage metacognition and curiosity perceptions in order to reinforce the benefits of our "KidsAsk" training on children's spontaneous QA behaviors (see Chapter 5). To do this, we design a novel training we call "KidsReflect" where we focus on helping correct children's misconceptions of curiosity and on giving them specific declarative and procedural metacognitive training to help them better control their curiosity during learning.

The training consisted of two parts: one that presented animated videos to give declarative knowledge about curiosity and its metacognition skills and their important role to make learning more efficient and enjoyable. And one that had hands-on sessions to procedural and experience-based knowledge about these concepts using a web platform we designed. During these practice sessions, children were instructed to use their curiosity—as shown in the videos— to learn novel information about some texts. They had the help of conversational agents that remind them on how to use their curiosity and metacognition to solve a problem, by following some specific learning steps and strategies.

Our results show that participants who had the "KidsReflect" training —compared to those who only had "KidsAsk"— ended up with more progress in terms of divergent QA performance both during the "KidsAsk" training (where they had an agent to help them formulate their questions) and during the offline tests where they had no external incentive or support to do it. These participants also had stronger progress in their metacognitive sensitivity (i.e. accuracy in judging their knowledge level and performance) and their curiosity perceptions.

While our results show stronger effects for children who had the full "KidsReflect" training, we still see rather positive results for those who only had the declarative part (i.e. the videos). Indeed, we see a significantly higher increase in their divergent QA measures, compared to the group who had no "KidsReflect" at all. These findings are exciting as they suggest that even simple methods such as videos screenings can be rather efficient in changing students' perceptions about curiosity and helping enhance it during learning. They also encourage EdTech designers to work on developing tools that reinforce these metacognitive skills as we see stronger effects for children who had the hands-on digital practice sessions.

Taken together, our studies suggest the efficiency of combining interventions that explain and practice the linguistic, metacognitive and social components of curiosity in enhancing children's spontaneous curiosity-driven learning strategies. Using new technologies such as LLMs seems to help facilitate the implementation of such interventions on the larger scale and for various learning activities.

6.2 FINDINGS, LIMITATIONS AND FUTURE DIRECTIONS

6.2.1 *Functional ingredients for an efficient training of divergent QA behaviors*

By assembling the various insights we gained from the studies conducted throughout this thesis, we manage to identify and empirically substantiate three key factors that seem to facilitate the training and promotion of divergent questioning skills during learning. See Figure 6.1 for a summary of these findings.

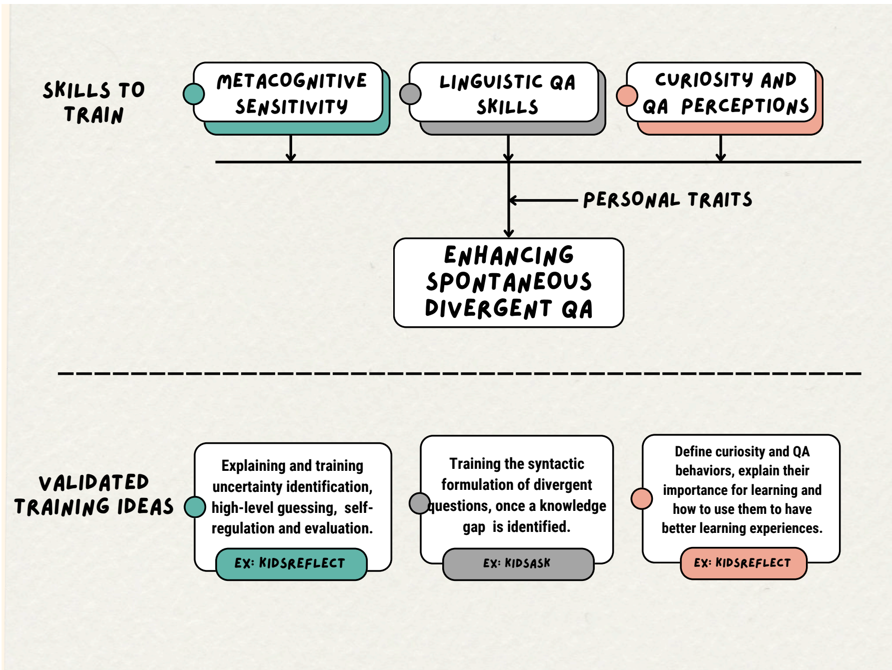


Figure 6.1: Functional ingredients for an efficient training and enhancement of divergent QA skills and instructional design principles ideas to support them.

LINGUISTIC QA SKILLS First, we see that children's performance in "KidsAsk", i.e. their ability to understand and benefit from a training that is designed to help them gain linguistic QA skills, seems to be a predictor of their subsequent offline and spontaneous divergent QA behaviors.

This is a rather expected result as we know that formulating divergent and high-level questions is not an easy task for children [69]. Indeed, besides relying on metacognitive aspects such as identifying a learning goal, generating divergent questions also requires advanced syntactic, linguistic and semantic abilities that children may not be fully comfortable with at this age [79, 80]. It is in this context that we designed "KidsAsk": it aims to take off the cognitive load of thinking about what knowledge gaps to pursue (by having an agent that makes these KGs explicit) and focuses more on how to transform the latter into well-formulated divergent questions. To further facilitate the training, the agent also gives linguistic cues in the form of high-level questioning words that children could use in order to avoid simple ones that usually lead to convergent questions [42].

These results are also in line with previous QA models such as in [169] suggesting the importance of language in learning in general and in curiosity-driven QA more particularly. Here, the importance to "finding the right words, giving shape to an idea and articulating what is meant" is highlighted. In their model, Ronfard et al. also state that question-asking can be divided into four components: (1) initiation, (2) formulation, (3) expression, and (4) response evaluation and follow-up [192]. Where formulation refers to the ability to phrase the question so that it can be understood.

METACOGNITIVE SKILLS Second, we also see that children's metacognitive sensitivity was a significant predictor of their spontaneous divergent QA behavior. This index has indeed increased after the "KidsReflect" intervention and was correlated to the understanding and performance during both parts of the "KidsReflect" training.

This result supports previous hypotheses suggesting the link between metacognitive skills and the stimulation of epistemic curiosity [55, 71, 142, 226]. But despite the abundant literature, this link was not often put into evidence in educational contexts. The work conducted during this thesis thus brings a new step towards reinforcing these assumptions in these settings. It also reinforces the view of curiosity as a metacognitive function that requires identifying one's informational needs and predicting learning goals in order to guide the exploratory behavior [43, 77, 91]. This is also in line with previous QA models such the one introduced by Ronfard et al. [192] where they suggest that the first step in the process of asking a question is *initiation*, i.e. the "realization that information is lacking and needed". This step is indeed inherently metacognitive as it requires self-evaluation and continuous monitoring to recognize missing information.

On a final note, it is worth mentioning that even though our "KidsReflect" training seems to have a rather small effect of children's metacognitive sensitivity (probably due to its short duration), it still shows significant progress for children that had the full intervention. This suggests that improving metacognition as a domain-general skill and transferring it to specific learning tasks is possible. This can be seen as a new result both concerning the age range we targeted (9 to 11 years old) and the methods we used. To the best of our knowledge, previous work aiming to train metacognition was targeted towards adult college students, as it is considered that metacognition requires high-level cultural and social resources that may not be available for younger students [20, 39]. Furthermore, these studies mostly had a domain-specific approach to train metacognitive accuracy by giving specific feedback about students' performance [38, 160]. They also used external incentives to motivate students' accurate self-reflection (e.g. extra credit) rather than intrinsic ones such as what we used during our work (e.g. understanding the role of accurate self-reflection for learning progress, etc).

PERCEPTIONS OF CURIOSITY Similar to metacognitive sensitivity, children's perceptions of curiosity have also predicted their spontaneous divergent QA, along with metacognition and syntactic QA skills. We also see that they have evolved over time after our interventions.

First, we find that children's initial perceptions of curiosity and QA behaviors in the classroom were generally rather negative, often associating it with stubbornness, stupidity, etc. This finding adds indeed to our motivation to implement a training addressing this challenge.

Second, we saw that children who benefited from the "KidsReflect" training, whether in its declarative-only or complete form, were able to enhance their perceptions, unlike those who didn't have the training. Furthermore, we find that this change is correlated to their understanding level of the videos presented during part I of where we gave conceptual knowledge about curiosity, its importance and how to control it using metacognition. These are exciting results as they show that it is possible to correct children's negative perceptions of curiosity, using the appropriate explanations and trainings, even via simple and easy-to-implement methods such as video screenings.

Finally, we also find that the perception of curiosity is a predictor for their willingness to spontaneously ask curiosity-driven questions, without the need for external help or incentives. These findings are in line with the idea that children's curiosity is not only intrinsic, but also develops through social perceptions and interactions [58]. For instance, in their model described above, Ronfard et al. also highlight an *expression* step in their QA model. This step refers indeed to the decision of whether it is 'worthwhile' to ask a question once formulated, both in terms of informational efficiency and social acceptability [192]. Furthermore, several empirical studies have found support for this idea in the educational settings. For example, in a study by Henderson and Moore [95], children's explorations were found to be influenced by the teachers' attitudes and behaviors (i.e. whether or not they were attentive, supportive and encouraging of exploration and question-asking). Interestingly, they also show that children whose curiosity behaviors were the most influenced by teachers, were those with the lowest curiosity scores as reported by their guardians. Similarly, Post et al. [183] found that students' fear of negative judgement from teachers and classmates led them to have a negative opinion about asking questions in the classroom and associating this behavior with negative elements such as being stubborn or stupid. In another study, Marx shows that the tables arrangement in the classroom can have an impact of children's question-asking fluency [158]: those who were not put in standard settings (facing the teacher, which usually establish a vertical relationship with the teacher, inspires fear of judgement, etc) were asking significantly more questions.

PERSONAL TRAITS Finally, we find that individual differences in curiosity trait (as reported by our participants' guardians) influenced their divergent QA behaviors, but not in all settings. Indeed, children who only had the "KidsAsk" training did not show any correlation between these two measures, except for the condition where the agent gave open cues leading to several possible questions. As we discussed above, this result is somehow expected given that the "closed" conditions, as opposed to the open one, might be providing training for the linguistic part of the question-asking rather than the curiosity one. Indeed, the cues given by the agent in the 'closed' cases may not match children's ZPD and therefore, they are not disposed to express their own curiosity [159].

However, for children who also had the "KidsReflect" training, we see that curiosity trait influences their final offline divergent QA fluency, even though they interacted with the incentive "closed" agent during "KidsAsk". This may lead us to suggest that with a reinforced conceptual and procedural knowledge of curiosity and metacognition, children were more able to see and benefit from "KidsAsk" as a way of transforming already-identified uncertainties into well-structured questions. They were then able to apply this procedure in other activities to serve their own curiosity by asking the questions that are of interest to them. This is also a rather expected result given that one of "KidsReflect" main objectives is to teach children how to become more aware and to control their own curiosity by explicitly thinking about their own knowledge and learning goals.

Together, these results underscore the relationship between trait curiosity and curiosity-driven behaviors such as ability to ask divergent questions, a finding also noted in previous studies [8]. They also align with the psychological concept of personality traits in general as dispositional qualities with multiple origins (genetic, environmental, etc.) that result from a positive interaction between the individual and their environment. In this sense, we can suggest that building educational interventions targeted to train curiosity can thus be a good strategy to address and mitigate inequalities around active and open-ended learning.

6.2.2 *Limitations and perspectives*

While our studies show rather positive results, our methods still present some limitations, therefore inspiring some promising future directions. These limitations concern, mainly, the lack of long-term assessment of the impact on children's curiosity and learning, the challenges around integrating our methods in formal learning settings and activities that could be led independently by teachers, the lack of personalization and responsiveness in our technical solutions and, finally, the challenges around the scaling-up of these methods in industrial EdTech solutions.

PEDAGOGICAL AND EXPERIMENTAL DESIGN LIMITATIONS First, as we mentioned in the previous sections, one of the main limitations of our first implementations of "KidsAsk" is the stringent nature of the agent's behaviors (i.e. its cues lead to specific and pre-defined questions, leaving children with very little choice over the questions they can generate).

However, we also see that this training, compared to the one with the 'open' strategy, was more beneficial for children on the linguistic level. Indeed, given that the incentive cues were proposed in the form of short sentences, they helped children gain insight on how to structure their questions and formulate them correctly.

As mentioned above, this linguistic dimension is important as it helps children be more comfortable with expressing their curiosity, through asking high-level questions to other agents [69]. In this sense, we think that "KidsAsk" with the incentive implementation might be seen more of a linguistic training rather than a curiosity one, but remains necessary for children to gain competence in how to practically and efficiently express their uncertainties, once identified. The skills of identification and evaluation of these uncertainties could be seen as other competencies that require different mechanisms for training, as we show with the "KidsReflect" intervention.

During the "KidsReflect" training, whether in its declarative-only or declarative + procedural form, children learn about the metacognitive skills essential for curiosity-driven learning through a series of sequential and ordered steps (i.e. self-reflect, *then* guess, *then* explore, etc). In other words, our training teaches children a specific procedure to be applied in a specific manner.

This approach can be problematic because it raises the question of whether children can deconstruct this procedure and understand its components individually. This is crucial, as the ability to understand and apply each skill independently is essential for adapting these latter to the specifics of each learning situation we face, thus ensuring effective transfer of curiosity-driven and MC-regulated learning strategies.

MEASUREMENT LIMITATIONS First, throughout the studies presented in this thesis, our analyses primarily focused on assessing the mean divergence score of the questions generated by participants. These questions were produced during specific tasks administered at various stages of our intervention and were assigned divergence scores based on subjective evaluations by two different raters.

While this method presents the advantage of being behavior-based rather self-report-based, it can still have some limitations due to its subjective nature. For example, when information is available but not understood by a child, they will generate a question about it, perceiving it as divergent. However, such questions are typically given a divergence score of 0, despite serving an epistemic function and addressing a specific knowledge gap from the child's perspective given that they pursue information that is already available.

Therefore, we think it is important to develop novel, objective measures for assessing divergent thinking and curiosity that do not rely on subjective ratings. One potential approach is the scoring method for divergent thinking tests proposed in [209]. In this method, participants complete the divergent thinking task and then select the two responses they consider their most divergent. Raters then evaluate these responses on a 5-point scale, enhancing the validity and reliability of the measure.

Second, for the metacognitive sensitivity measure we take during the "KidsReflect" training, it is arguable that this measure (i.e. how good the individual is in distinguishing good from bad performances) is not sufficient to assess individuals' metacognitive skills, and especially those related to curiosity.

Indeed, the regulation of cognition is also an important metacognitive component to curiosity as it can facilitate adopting the relevant and efficient strategies to search for information given each specific learning context. Similarly, metacognitive monitoring can be responsible for fostering epistemic vigilance and maintaining curiosity-driven learning cycles until a satisfying answer is found [43]. Therefore, it seems important to assess this metacognitive regulatory and monitoring skills in future implementations. This can be assessed using some behavioral measures such as frequency of errors detection and self-correction [134, 178], or using self-report-based measures such as the MSLQ questionnaire that captures metacognitive monitoring and self-regulation [181].

Furthermore, it is important to say that our measure of metacognitive sensitivity using a type 2 ROC analysis, can still have its limits. Indeed, while this measure presents the advantage of being bias-free compared to self-report or correlation measures (in theory, it is not influenced by the subjects propensity to report a specific state of confidence [63]), it can still be influenced by other factors such as task performance [70]. This means that a change in an individual's metacognitive sensitivity may depend on a change in their task performance [99]. Concretely, this calls for the need to control for the task performance prior to calculating the index, which is challenging.

Finally, we highlight that during the "KidsReflect" training, we did not measure domain-knowledge learning progress due to experimental design constraints (need for additional sessions and time. This is an important limitation of our study, considering that our ultimate goal is to evaluate the efficiency of training curiosity-driven and MC-regulated learning strategies as a means to enhance learning.

This limitation underscores the need for longer and more ecologically valid implementations of "KidsAsk" and "KidsReflect." Such extended implementations could allow us to investigate the impact on domain-knowledge learning progress more thoroughly.

ECOLOGICAL LIMITATIONS As mentioned earlier, a clear next step for us is to explore how we can motivate pedagogical teams and industrial EdTech designers to adopt the methods introduced in this thesis for longer-term use in everyday classroom activities.

Although our studies made progress in designing interventions conducted in classroom settings, they were still heavily controlled by research methods. Indeed, during all our interventions, we followed specific experimental procedures that do not necessarily reflect real classroom environments, e.g. work individually on tablets, etc. This suggests that our results might be influenced by these somewhat artificial conditions and could differ when applied to formal classroom activities. It is also to be noted that the presence of the research team was always mandatory when conducting our training, leaving teachers with little opportunities to familiarize themselves with the methods and manage the interventions independently.

In this context, we started working towards addressing these challenges, following three principle axes:

- **Formalize feedback from teachers whose students tested our tools:** We aim to obtain a quantitative evaluation of the usability, usefulness, and acceptability of the methods and tools we proposed during our interventions. It is important to note that, although researchers conducted all the training sessions, teachers were always present, observing each session and testing the tools before their students used them.

To gather this data, we use theoretical frameworks such as those described in [67, 188] to define specific indicators for surveying teachers. These indicators fall into three categories: 1) indicators for usability: measure the ease and comfort of use, flexibility and adjustability, and workload (i.e., resources needed to prepare and/or work with the tools). 2) Indicators for usefulness: to measure the relevance of the pedagogical objectives and nature of tasks, relevance of task and session duration and sequencing, comparison to other tools or methods, evidence of students' motivation, attention, and learning progress. And finally, 3) indicators for the acceptability of the tools: to measure their compatibility with the teachers' ethics and values, schedule, organization and teaching methods, and their support of teachers' professional interests and development.

By gathering this information, we adopt a continuous design process to ensure the usefulness of our methods on both didactic and pedagogical levels.

- **Train teachers to carry out the training independently:** Our first direction is to minimize the necessity of the research team's presence during the trainings. Indeed, conducting interventions in a controlled environment can differ significantly from mainstream school settings, potentially leading to biased results. In this context, a new PhD project within the team is starting to collaborate with teachers to help them become self-sufficient in delivering the training.

  The goal is to educate teachers in cognitive science—specifically, the theories of curiosity and metacognition that underpin our interventions—and in technical aspects, such as operating the software and assisting children in its use. This training will enable teachers to take full ownership of the methods and apply them autonomously, aligning with their teaching methods.

  This is indeed a crucial dimension as introducing the teacher can increase the efficiency of the training and make it more suitable for children's interests and ZPD [159, 179, 225].

- **Co-create formal curiosity- and metacognition-supporting learning activities with teachers and EdTech industries:** As previously mentioned, our training methods were implemented in informal and short learning tasks. While this approach allowed us to implement controlled interventions and scientifically validate our methods' efficiency, it also poses a limitation for scaling them up. Our next goal is thus to advance towards more general implementations.

In this context, our first approach is to work on integrating our instructional design principles into standard and formal learning subjects and activities that children encounter during their education (e.g., mathematics, history, language acquisition, etc). There are two main possibilities in this context: 1) Work with teachers to co-create offline and standard activities using our instructional design principles. These activities can be directly implemented in classrooms to meet pedagogical and curriculum needs. And 2) Collaborate with educational technology designers and pedagogical engineers to adapt our approaches to the content they provide, that supports formal educational and pedagogical goals. Leveraging platforms from our industrial partner, EvidenceB, which are used by various schools across the country, can help achieve this goal.

In the context of industrial up-scaling, we began exploring ways to assess the usefulness of the general methods developed in this thesis within marketed educational technologies and Intelligent Tutoring Systems (ITS). We collaborated with the young EdTech startup EvidenceB to implement and test these investigations.

EvidenceB develops a K-12 e-learning platform that personalizes educational content aiming to maximize both learning efficiency and intrinsic motivation. Their personalization approach uses an online reinforcement learning algorithm inspired by the LPT [176]. The algorithm predicts and proposes specific activities from a large pool of exercises to help students achieve optimal learning progress for the knowledge component they are working on [46].

However, several field tests revealed that students reported low motivation when using the platform. One hypothesis is that this is due to students' lack of metacognitive skills to monitor and recognize their learning progress, a key factor in supporting IM and engagement according to LPT.

Inspired by the literature and the methods explored in this thesis, we worked on facing this problem by adding support features to the platform in order to help students monitor their learning progress. These features were in the form of learning dashboards that students can view throughout their training sessions. They offer an 'overview' to show progress over time and a 'details' view to show knowledge levels for each pedagogical goal. The aim is to provide tools that support procedural evaluative and monitoring metacognition, which we hypothesize will help students recognize their learning progress. This recognition is necessary to form the positive feedback loop with IM, as stated in LPT. Evaluations of this tool's usefulness and its impact on learning and motivation will be conducted in the coming months as the product is currently being implemented in numerous French public schools.

Future projects will focus on large-scale implementation of the metacognitive scaffolding methods developed in "KidsAsk" and "KidsReflect" (Chapter 3 and Chapter 5). Specifically, we aim to use large language models (LLMs) to generate personalized cognitive and metacognitive hints that can guide students during their learning tasks. Similar to the work presented in this thesis, these hints will have a specific structure to help students become aware of their learning strategies and performance and support them in formulating questions when needed.

In the long term, we also aim to personalize the frequency of this support based on students' emotional and cognitive states. Ultimately, we hope students will adapt these self-regulated metacognitive learning strategies and use them independently without the need for external support. This approach can promote student engagement without relying on extrinsic rewards, such as intense gamification or monetary incentives, which we often see in EdTech software and that can have mixed impacts on learning.

Finally, the positive results demonstrated in our work regarding GPT-3's efficiency in generating pedagogical content (Chapter 4) have inspired new directions for EvidenceB's ITS. Indeed, Several projects are exploring methods to use advanced LLMs such as GPT or Mistral to train smaller models on tasks of pedagogical content generation, while respecting specific cognitive and pedagogical goals. This approach aims to reduce the production costs, as currently, all content is manually created by experts in cognitive and learning sciences, which is a highly costly process.

In conclusion, we think that the work conducted during this thesis can provide EdTech designers with instructional design principles to implement new features and motivational mechanisms. These can maximize the efficiency of their products in terms of both learning outcomes and key learning predictors, such as motivation, curiosity, and metacognition.

TECHNICAL LIMITATIONS Finally, we address the technical limitations of the platforms used in our "KidsAsk" and "KidsReflect" interventions.

The primary limitation is the lack of feedback provided to participants regarding their performances. Our platforms currently lack NLP methods that could offer children real-time assessments of the quality of their performance, such as the divergence level, complexity, and pedagogical value of their questions. Additionally, we do not provide feedback on the metacognitive steps leading to question generation, such as the relevance of an identified uncertainty, the quality of a guess, or the accuracy of evaluating the information resulting from a question.

For all these activities, we only assess the response quality offline and manually, once the experiment is over. This is a significant limitation of our study, as feedback is crucial when learning new tasks, especially novel ones that children are not used to doing like those we propose. We hypothesize that receiving informative feedback about their performance would help children develop a better perception of their competence in these tasks (identifying uncertainties, asking relevant questions, etc.), and thereby motivate them to adopt and improve such behaviors [176].

A promising future direction is thus to investigate novel ways to include real-time evaluations of children's performance in metacognitive and curiosity-driven tasks. Advances in NLP and the emergence of LLMs more particularly can support such investigations [166, 231]. Integrating these technologies could provide immediate, informative feedback that could enhance children's learning experiences and outcomes.

In this context, and through a collaboration with NLP scientists, we took a first step towards exploring the use of the GPT-3 LLM to automate the annotation of the quality of curiosity-driven questions [229][1]. We looked at two dimensions: the question complexity and its syntactic structure. These two dimensions are taken from the annotation grid we used to manually assess children's question during our studies, see Appendix B.4). The complexity level looks at if the answer is a simple fact (e.g., "How big is a dinosaur?") or requires explaining a mechanism, a relationship, etc.(e.g., "Why were dinosaurs so big?"); it gives a binary score. The syntactic structure has a score that goes from 1 to 4 points: 1) 'closed' or declarative questions (e.g., "Dinosaurs were big?"), 2) questions with questioning words in the middle of the sentence (e.g., "The dinosaurs were how big?"), 3) questions without an interrogative formulation (e.g., "Why the dinosaurs are big?"), and 4) questions with a questioning word at the beginning of the sentence that has interrogative syntax (e.g., "Why are dinosaurs big?").

Using all questions generated by children during all of our studies reported in Chapter 3 and Chapter 4, we collected a dataset with a total of 668 questions in French [2]. We then coded each question on the dimension of question complexity and syntactic structure, following the grid described above. Additionally, the dataset and codebook have never been published online and are thus unseen by LLMs.

1 The text in this following section is taken from the publication in Ziang Xiao et al. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding." In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023, pp. 75–78

2 For question complexity, we first used GPT-3 to translate questions into English. For the syntactic structure, we kept questions in French to preserve its syntactic structure.

With a team of NLP scientists, this coding scheme was then run using GPT-3 (davinci-text-002) with a temperature of 0.0 during the prompting process. We choose this setting because it was the most advanced version of GPT-3 that was publicly available when we conducted the experiments and a temperature of 0.0 to guarantee the reproducibility of this study.

Two design dimensions of the prompt were explored:

- codebook-centered vs. Example-centered: This dimension regards the structure of a prompt. In the codebook-centered prompt, the prompt was designed similar to how we read a codebook. The prompt follows the structure of [Code/ Description/ Examples]. For example, Code: HIGH; Description: the answer to this question is not a simple fact but requires explaining a mechanism, a relationship, etc.; Examples: Why were dinosaurs so big? The example-centered approach is inspired by the in-context learning in LLM works where the prompt explains the rationale behind each example [146]. For example, "Why were dinosaurs so big?" is an example of "HIGH" because the answer to this question is not a simple fact but requires explaining a mechanism, a relationship, etc. The code, examples, and descriptions are the same for both designs.

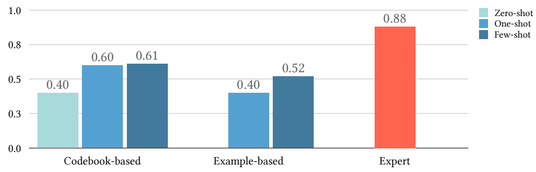
- Zero-shot vs. One-shot vs. Few-shots: Since recent work showed conflicting results on the number of examples in a prompt [146], we explored different prompt settings. In the Zero-shot setting, we give only Code and Description in Codebook-centered prompts [3]. For the One-shot setting, we provided only one example for each code. And for the few-shot setting, we provided five examples for each code.

For all prompt variants, we included an identity modifier, "I am a developmental psychologist who has expertise in linguistics.'

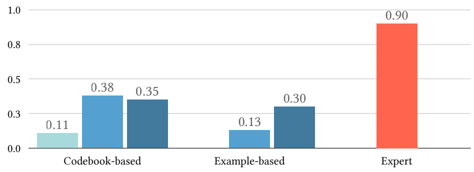
We measured the performance of our GPT-3 based approach with Cohen's Kappa. Cohen's Kappa measures inter-rater reliability, which indicates how two coders agree with each other. We computed two sets of Cohen's Kappa, between GPT-3 with the expert's final coding results and between two experts who originally coded the dataset using the *same* codebook.

The results suggest that it is feasible to use GPT-3 with an expert-developed codebook for deductive coding. When analyzing curiosity-driven questions, the GPT-3-based approach achieved fair (Syntactic Structure: Cohen's $\kappa = 0.38$) to substantial (Question complexity: Cohen's $\kappa = 0.61$) agreement with expert rating, see Figure 6.2. However, there is a gap between experts' agreement with our approach and the agreement among experts (Question complexity: Cohen's $\kappa = 0.88$;

3 Since the example-centered approach requires at least one example, we did not have the zero-shot setting for the example-centered approach



(a) Cohen's κ for Question Complexity



(b) Cohen's κ for Syntactic Structure

Figure 6.2: The Cohen's κ between GPT-3 and experts shows substantial agreement in Question Complexity coding and fair agreement in Syntactic Structure coding. In general, Codebook-centered prompts with examples achieves the highest agreement [229].

Syntactic Structure: Cohen's $\kappa = 0.90$). The different prompt designs were also examined. The codebook-centered design was found to perform better than the example-centered designs, see Figure 6.2. And examples play an important role. The largest performance gain when shifting was seen from a zero-shot to a one-shot setting. However, the performance between one-shot and few-shot settings did not differ much.

These preliminary findings indicate the feasibility of using LLMs for qualitative analysis of children's curiosity-driven questions. By combining GPT-3 and a codebook, the LLM-based approach achieved fair to substantial agreement with experts. Considering the accessibility and flexibility of LLMs and the new models such as GPT-4 that show stronger performance in this task (e.g. [166]), we believe this approach has the potential to effectively help analyze children's questions and give real-time feedback about their quality. A next step will also be to develop a similar mixed method to evaluate children's metacognitive steps during learning tasks.

6.3 OPEN QUESTION: WHAT ROLE FOR CURIOSITY-DRIVEN LEARNING IN A GENERATIVE AI AGE?

With the current unprecedented rise of Generative Artificial Intelligence (GAI) and especially LLMs, there is widespread enthusiasm about their potential to address various challenges in the educational sector. This excitement particularly stems from the possibilities GAI offers, such as facilitating the implementation of personalized tutors for every learner. These personalized tutors can foster intrinsic motivation, engagement, and curiosity, all of which are positive predictors for learning [122].

Within this general enthusiasm, and with the release of new tools that are increasingly adept at resolving complex educational tasks with minimal effort from learners, we pose an important question: Is it still relevant to emphasize the importance of question-asking in an age where LLMs can provide answers without requiring well-structured questions?

6.3.1 *Curiosity-driven learning is still crucial, but is endangered by GAI*

We can argue that curiosity-driven learning, metacognition and active learning strategies are even more important when students are interacting GAI. We present a two-step argument to explain this.

1) While associated with huge capabilities even for the most complex tasks, GAI tools are still subjects to several fails (ambiguity, cognitive bias, etc; these fails seem also to be harder to detect for humans compared to human-generated misinformation [41]. For instance, research has shown that even university students struggle to identify inaccuracies in ChatGPT's answers and may use them in exams without proper analysis or correction[100]. Such results are not surprising when considering children's learning mechanisms from an early age [124]. Young learners tend to track the knowledgeability of the agents they interact with and use this signal to form their own beliefs and shape their information-seeking behaviors [94, 174, 199]. However, if students have underdeveloped metacognitive skills and overestimate the knowledgeability of LLM-based agents, they can easily adopt incorrect and/or biased information and learning strategies. Once transferred, this misinformation and these biases can be difficult to correct [219].

2) This hindered metacognition, control of learning and over-estimation of LLM-based agents knowledgeability can come from two crucial GAI design principles:

1. GAI can challenge critical self-reflective skills: a key property missing in LLMs' behavior is uncertainty signaling. Indeed, LLM-based systems consistently exhibit confidence in their outputs, even when they are incorrect. This constant affirmation means that students have limited opportunities to critically evaluate the quality of the information they receive. This can, in turn, impair their metacognitive abilities to reflect on their own knowledge state and assess their progress toward achieving specific tasks. As suggested in several theories of curiosity, overestimating one's own knowledge and lacking evaluative and monitoring metacognitive skills can significantly hinder individuals from engaging in exploratory and active information-seeking behaviors [147].

2. GAI can challenge the ability to actively work for information: students can become overly reliant on receiving information with minimal effort. This reliance may lead them to believe that solving any problem is easily accessible without the need for specific prior knowledge [122]. Such passiveness can undermine intrinsic motivation to engage in investigations and active information-seeking behaviors. Moreover, the apparent ease and convenience of obtaining information from LLMs could result in negative learning control [15, 28], where students begin to associate cognitive success with these powerful tools rather than with active learning strategies such as question-asking and exploration.

We therefore emphasize the importance of curiosity-driven learning strategies and metacognition for promoting informed use of GAI in education. Without these skills, GAI-based learning tutors might actually work against the pedagogical objectives established by educational research. Empirical research is starting to support this idea. For instance, a randomized large-scale study with students from several continents shows that while using GPT-4 in coding classes may have helped student learning, it also posed potential harm to their long-term engagement [173].

6.3.2 *Supporting curiosity-driven learning and metacognition in a generative AI era*

We believe that the challenges we state above are mainly due to the lack of a pedagogical stance in GAI tools, i.e. the lack of alignment between pedagogical goals and GAI design principles and purposes. Indeed, LLMs are not designed —and do not behave in a way— to carry out pedagogical goals and support users' learning and activeness. On the contrary, they are designed to give away answers with the minimal effort possible needed from users. Even with training methods such as RLHF that are put in place to align LLMs with human goals, it is still not clear what is the pedagogical intent (if any) behind the training of these models. Indeed, RLHF involves human evaluators who can be of various backgrounds, opinions, goals and expertise.

In this context, and as we illustrate in Figure 6.3, we suggest that facing these challenges can be initiated by different stakeholders and on different levels:

- **Role of the AI community on the training level:** consists of refining the training policies and data curating methods used for the models, in order to achieve a pedagogical method/ goals-grounding. Concretely, this could be done by training the model on more educational and learning principles data, recruiting teachers as annotators during the RLHF process, etc.

- **Role of educators/ EdTech designers on the implementation level:** consists of making informed decisions when using LLMs to design educational activities (e.g. by ensuring enough control, pedagogically-oriented interactions, etc).
- **Role of researchers and the educational sector on the usage level:** consists of preparing students to be aware of GAI's challenges and be able to use it in an informed and positive way. It relies on supporting them to develop the relevant skills that could help them stay active and critical when using this technology.

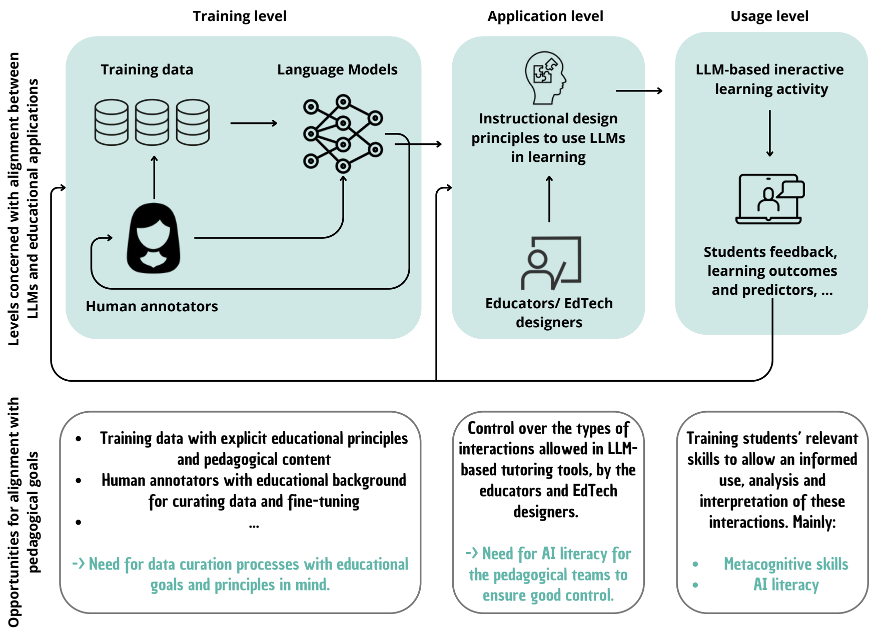


Figure 6.3: Life-cycle of LLMs' usage in educational applications and opportunities for alignment with pedagogical goals [4].

In focusing on the third point above, our suggestion is that students need to be equipped with the relevant skills that can allow them to be in control over their interactions with LLMs. This control is also key to nurture their curiosity and help them stay engaged in the process.

We hypothesize that there two type of skills essential for fostering this: 1) GAI literacy-related: as already discussed, on of the biggest challenges that can hinder children's exploratory and critical reflective behaviors is the misleading over-hype around GAI's agency and knowledgeability. To have a more realistic representation of these tools, we thus think that it is important for students to see their real potential, strengths and weaknesses and understand the motivation and rationale behind their design. We also think that the educational teams (educators, pedagogues, etc) should have access to GAI literacy in order to be able to propose GAI-based activities that still follow their educational goals and keep their students engaged. And

2) metacognitive-based: if we want students to be more at ease with asking for further information when they feel surprised or/and uncertain about GAI's behavior, supporting them to develop evaluative and monitoring metacognitive skills becomes essential. Indeed, as seen in this thesis, these are important skills that can support critical-thinking mechanisms and the ability to identify knowledge discrepancies when interacting with other agents. Training the formulation of efficient questions, following specific informational need, is also a crucial skill.

6.3.3 *Study: "Hey ChatGPT, explain this to me!" On students' efficiency in using GAI to solve learning tasks*

Before moving to designing pedagogical interventions that focus on training these crucial skills, we realize that there is a gap in literature. Indeed, to the best of our knowledge, very little empirical work has explored how young learners interact with these tools: how they use them to solve educational tasks, the type of inquiries they use, and the effect on their learning strategies, etc.

In this context, we designed a new study aiming to bridge this gap. The study aims to understand: 1) students' efficiency in prompting a GAI tool—ChatGPT— in terms of specificity of their queries and quality of questions when searching for information to solve specific educational tasks. 2) Their efficiency in finding accurate answers using ChatGPT and the effort it takes them to do so. And finally, 3) The relationship between these two indicators and students' perception of ChatGPT (knowledge of how to use it, its limits, etc) as well as their metacognitive and QA abilities.

To do this we recruit 72 middle-school students in the South-west region of France, aged between 14 and 16. During a 1-hour session, they first start by completing three questionnaires. The first one concerns their perception of LLMs in general an their use in education more specifically.The questionnaire we give is an adapted French version of the one proposed by Bernabei et al. [26] for college students. In order to be adapted for the age range we are studying, our version contains less items while preserving the original dimensions and their psychometric quality. It contains scales for: attitude towards LLMs, trust, social influence, fairness & ethics, usefulness & performance expectancy and finally, effort & ease of use.

The second questionnaire our students fill is the Junior Metacognitive Awareness Inventory, an 18-item questionnaire developed in [214] with two dimensions: knowledge of cognition and regulation of cognition. The two questionnaires are available in Appendix A.8 and A.3.

The final test concerns their question-asking fluency. For this, we use the same process as for our previous studies, described in Appendix A.10. The idea is to collect students' questions for text and evaluate the quality of these latter (structure, specificity, etc).

After filling out the three questionnaires, participants move to do a learning task. The task consists of 6 science-related exercises that students need to answer with the help of ChatGPT. Students are told that their goal is to use this tool to solve the tasks the most accurately and quickly possible.

For each exercise, students see a suggestion for a question they could ask ChatGPT in order to find the answer they're looking for. These suggestions are randomly assigned: for half of the tasks, they contain all necessary elements of context so that ChatGPT can generate an accurate answer directly. For the other half, they miss an element of context which will lead ChatGPT to either be unable to provide an answer or provide a very generic one. See Figure 6.4 for an example of a task with a 'correct' prompt and an 'incorrect' one. Participants can thus choose to either use our suggestion of prompts or not, to solve the task. They can ask as many questions to ChatGPT as they want, until they find an answer they are satisfied with and submit it.

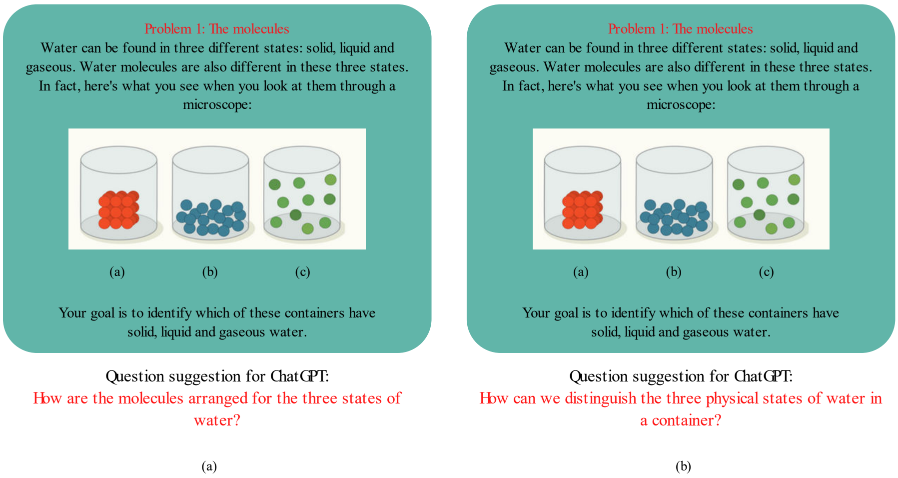


Figure 6.4: Example of a learning task proposed. (a) the task is accompanied with a 'correct' prompt for ChatGPT that will directly lead to finding the answer of the exercise. (b) the task is accompanied with an 'incorrect' prompt that is missing elements of the task context, leading ChatGPT to a generic answer that do not align with the exercise's goal.

With this design, our aim is to investigate students' ability to distinguish between efficient and inefficient prompts, i.e. their awareness of the prerequisites needed when interacting with an LLM-based agent in order for this latter to be an efficient learning companion (need to provide specific context, clear inquiry, etc). When participants choose to formulate their own questions to ChatGPT, we will also assess the quality of these latter using manual annotations. Linking these measures to students' perception of LLMs, their metacognitive abilities and their QA skills can indeed be an important observation for us as it will motivate more focus on training these skills in order to allow an informed use of LLMs in educational contexts.

Finally, we also analyze the learning measure, i.e. the ability to solve the exercises accurately using ChatGPT and the effort associated with it. We focus on the number of interactions needed with ChatGPT to reach a correct answer, and whether or not students reformulate/ summarize ChatGPT's output to formulate their own answers. Once again, linking these measures to students' perception of LLMs and their metacognitive abilities can give us an insight to the prerequisites needed to have an efficient use of these tools to solve learning tasks [4].

This study is a preliminary step in a broader and more general question surrounding the long-term impact of GAI on the educational landscape, the learning and teaching strategies and efficiency. Indeed, as GAI tools are becoming increasingly accessible (e.g. offering simple interfaces, free access, requiring no specific skills to be used), they are likely to become students' go-to source of information to complete learning tasks.

This raises several questions about current educational practices. For example, how relevant is it to propose memory-based activities for evaluating students' understanding of a concept? More generally, what types of pedagogical goals should educators focus on in the age of GAI? How can we support students' expression of their individual differences in schools when information is readily accessible? These are crucial questions to address as we integrate GAI into educational settings.

It is therefore crucial to develop theoretical frameworks to analyze and regulate educational practices and goals in the GAI era, both for instruction and assessment. For example, emerging ideas suggest shifting towards activities that emphasize higher-level competencies, such as applying knowledge in novel situations, analysis, and creation.

Designing these frameworks presents a complex challenge, as it requires balancing ethical considerations with pedagogical needs. Aligning educational policies with the rise of GAI must also address inequalities in access to these tools. Therefore, it is crucial to foster collaboration between AI ethicists and educational scientists to develop precise, fair, and equitable frameworks.

In general, we We believe that future research should focus on interdisciplinary efforts to synergize work around the alignment of GAI

4 This is a work-in-progress study, the results and data analysis are not yet available. It is in the context of a collaboration with Pr. Celeste Kidd, at UCB.

to pedagogical goals and the fair re-thinking of educational practices to accommodate the role GAI is increasingly playing in society.

Part V

APPENDICES

A QUESTIONNAIRES AND OFFLINE TESTS

A.1 CURIOSITY TRAIT

This questionnaire is developed in [143]. The items can be answered by "Almost never", "Sometimes", "Often", "Almost all the time".

- My child has fun learning new themes/subjects.
- When presented with a difficult problem, my child focuses all his attention on how to solve it.
- My child is attracted to new things in their environment.
- My child devotes considerable effort to trying to discover confusing or unclear things.
- My child enjoys talking about topics that are new to them.
- My child is confused they don't understand something, and strives to make sense of it.
- My child shows visible enthusiasm when the discover something new.
- My child will work long and hard to solve a problem because they want to know the answer.
- When my child learns something, they ask lots of questions about it.
- My child examines things by turning them around or looking at them from all sides.

A.2 PERCEPTION OF CURIOSITY

The items of this questionnaire are answered with a 4-point Likert scale. 1 means "I agree with the statement with a very small amount" and 4: "I agree with the statement with a very large amount". The questionnaire is developed in [183].

- To be curious is to want to know how someone learned a secret.
- To be curious is to want to know how someone learned a rumor.
- To be curious is to want to know how the human body works.
- To be curious is to want to know how a car works.

- To be curious is to want to know how computers were made.
- To be curious is to want to know how birds fly.
- To be curious is to want to know how mathematics was invented.
- I love asking myself questions about what I've seen in class.
- It's very important for me to ask interesting questions at school to learn more about what's around me.
- It's very important for me to ask questions in class to learn more about different things.
- I love asking questions about lots of things in class.
- I think that people who want to know a lot of things are important for the country.
- I think that people who ask good questions have a big impact on society.
- I think that people who ask interesting questions are very important for society.
- I think my classmates are stubborn when they always want to know everything in class.
- I think my classmates are annoying when they ask intelligent questions in class.
- I think people who ask lots of questions look stupid.
- I'm afraid my classmates will think I'm a nerd if I ask too many intelligent questions.
- I'm afraid my classmates will think I'm stupid if I want to know more about what we're learning in class.
- I find it scary to show that I want to know more about a subject in class.
- I'm really good at asking intelligent questions in class.
- I'm really good at asking new questions about different topics in my lessons at school.
- I think I'm really good at finding out new things at school.

A.3 METACOGNITIVE AWARENESS

We use the Junior Metacognitive awareness Inventory developed in [214] to assess participants' metacognition. The questionnaire contains items about knowledge and regulation of cognition. The items are answered with a 5-Likert scale. 1 means "Never", and 5 "All the time".

- I know when I understand something.
- I can force myself to learn when I need to.
- I try to re-use revision methods or strategies that have worked for me before.
- I know what teachers expect of me.
- I learn best when I already know something about the subject in question.
- To help me learn something, I make diagrams, drawings or graphs.
- When I've finished my homework, I check that I've retained what I wanted to learn.
- I think of several ways of solving a problem, then choose the best one.
- I think about what I need to learn before I start working.
- When I learn something new, I question the effectiveness of my learning strategies.
- I really pay attention to important information.
- I learn more when the subject interests me.
- I use my intellectual strengths to compensate for my weaknesses.
- I use different learning strategies depending on the task at hand.
- I regularly check that I'm achieving the goals I've set myself for my work.
- I sometimes use learning strategies automatically, without thinking.
- I ask myself if there's an easier way of doing things after I've completed a task.
- I set specific goals before starting a task.

A.4 GENERAL MOTIVATION

The general motivation scale developed in [47] has six items that are answered with a 7-point Likert scale. 1 point means "i don't at all agree with the statement" and 7 "I hugely agree with the statement".

- I enjoyed using the "KidsAsk" website
- I'd like to stay after school one day to use "KidsAsk".
- I think the "KidsAsk" site is useful to learn new things.
- I would advise a friend to try the "KidsAsk" website.
- I think I did well on the "KidsAsk" website".
- I would like the activities of "KidsAsk" to be harder next time.

These items are followed with two other items where children had to fill in the blanks:

- I found the "KidsAsk" site to be ... than my favorite video game.
- I found the "KidsAsk" site to be ... than my favorite subject at school.

The blanks can be filled with: a lot less cool, a bit less cool, almost as cool, as cool, a bit cooler, a lot cooler.

A.5 TYPES OF MOTIVATION

To understand types of motivation experienced during the tasks, we use Vallerand's scale [223] that distinguishes three types of motivation during learning: intrinsic, extrinsic and amotivation.

We asked children: "Why did you use the KidsAsk website, what motivated you to do so?". All items are Yes/No items:

- Because I'd be ashamed if I didn't succeed.
- Because I like to pass a test.
- Because it makes me happy when I answer one of the activities correctly.
- To have a gift.
- Because I'm happy when I learn a lot of new things that I didn't know before.
- Because I feel good when I use the site.
- Because I'm happy when I use it.

- I don't know why.
- I do what I'm told.
- To show that I'm clever.
- To get congratulations.
- Because I think that it is useful to learn about new things.
- Because I'm never bored when I use the site.
- Because I like learning new things.
- Because it's good for me to succeed in these activities.
- Maybe I'll find out later.
- Because I'm learning lots of things that interest me.
- Because I've always had good grades and I want to keep them up.
- To get a good grade.
- Because it puts me in a good mood when I succeed. Because I'm happy when I give 100% for an activity.

A.6 TASK LOAD

We used the Nasa-Tlx questionnaire to assess the load of the tasks we gave children. The questionnaire is developed in [92]. The six items are answered with a 20-point Likert scale. 1 corresponds to "very low" and 20 to "very high".

- How mentally tiring was the task?
- How physically tiring was the task?
- How rushed or hurried was the pace of the task?
- How well did you accomplish what you were asked to do during the task?
- How hard did you work to reach your performance level?
- How worried, discouraged, irritated, stressed and bored were you?

A.7 EXPOSURE TO DIGITAL TOOLS

We gave parents of participants a questionnaire to assess the exposure of children to digital tools in general. The questionnaire had 8 items that can be answered with "Not at all true", "Not true", "Neutral", "True" or "Very true":

- My child has interacted with a virtual robot before.
- My child has watched movies, TV series or other media containing virtual robots.
- My child is interested in virtual robots.
- My child would like to learn things with the help of a robot.
- My child would trust a virtual robot as a tutor.
- My child would follow the instructions of a virtual robot.
- My child would like to be friends with a virtual robot.
- My child would enjoy talking to a virtual robot.

The questionnaire had also the following questions:

- How many hours a week does your child spend on a phone, tablet or laptop? — could be answered by "0 to 5", "5 to 7", "7 to 10", "10 to 15" or "15 to 20".
- At home, how much exposure does your child have to different technologies (e.g. Google Home, smart home appliances) ? — could be answered with "None", "low", "average", "moderate" or "high".

A.8 PERCEPTION OF LLMS IN EDUCATION

We use a shorter version of the questionnaire developed in [26] to assess participants' perception of LLMs in order to make it more adapted to the age range we work with. We keep the same dimensions of the original questionnaire and calibrate the number of items per dimension to keep its psychometric quality.

Items are answered with a 4-point Liker scale. 1 means "I dot not at all agree" and 4 means "I strongly agree".

- I keep up to date with the latest developments in artificial intelligence.
- I've already used ChatGPT.
- I know ChatGPT's strengths.

- I know ChatGPT's limitations.
- I know how to use ChatGPT for school tasks.
- ChatGPT can make me more confident about doing schoolwork.
- ChatGPT answers are reliable.
- ChatGPT's answers are accurate.
- ChatGPT's answers are understandable.
- ChatGPT's answers are up-to-date.
- I plan to use ChatGPT because people around me use it.
- I plan to use ChatGPT to stay informed.
- Using ChatGPT can help me reduce my learning time and therefore do better in my exams.
- I don't see a problem with using ChatGPT to do schoolwork.
- ChatGPT can be used to spread misleading or false information.
- It is important for me to ensure the confidentiality of my data before using ChatGPT.
- The use of ChatGPT is bound to become widespread in the school environment.
- Using the results provided by ChatGPT can simplify the completion of school tasks.
- Using the results provided by ChatGPT will help me complete school tasks faster.
- Using the results provided by ChatGPT can help me get better grades at school.
- Using ChatGPT can motivate learning, as it allows me to work in a fun and stimulating environment.
- ChatGPT answers are directly usable without the need for change.
- Using ChatGPT to do schoolwork requires more effort than what I'm used to.
- Using ChatGPT to do homework takes more time than usual.

A.9 PROCESS FOR THE READING FLUENCY ASSESSMENT

For the reading fluency test, we use the standardized process developed in [138]. During this test, participants, one by one, are asked to join the experimenter in a room alone. They then read the following text in French:

"Sous la mousse ou sous le toit, dans les haies vives ou le chêne fourchu, le printemps a mis ses nids. Le printemps a nids au bois. Annie amie, du renouveau, c'est le doux temps. Amie Annie, au bois joli gamine le pinson. Dans les buis, gîte une biche, au bois chantant. Annie ! Annie ! Au doigt joli, une eglantine laisse du sang: au bout du temps des fééries viendra l'ennui. L'alouette fait ses jeux, alouette fait un noeud avec un rien de paille. L'hirondeau piaille sous la pente des bardeaux et,vif et gai, le geai, sur l'écaille argentée du bouleau, promène un brin d'osier. Au verger, dans le soleil matinal, goutte une pompe dégelée. On voit un bec luisant qui trille éperdument des notes claires et, dans les pampres d'or que suspend la grille antique, on surprend des rixes de moineaux. Au potager s'alignent les cordeaux; l'if est triste à l'horizon et lourd et lent l'envol des corbeaux. Un lac étire ses calmes rives et, quand le soir descend, le miroir de ses eaux reflète les poisons des brignoles perfides. Et, quand descend le soir, quand joue la pourpre du couchant, le ciel rougit ses eaux. Dans la moire de l'eau danse l'ombre d'un ècueil. Tout est cris! Tout est bruits! Une amarre est dècochèe . . . une barque est arrimèe . . . des matelots jettent leurs cassettes sur le rivage . . . Tout est cris! Tout est bruits! Au clair de la lune mon ami Pierrot . . . Au clair de la lune mon amie Annie. . . Au clair de la lune mon ami Pierrot, prête-moi la plume pour écrire un mot."

We then calculate the reading fluency score as the number of words spelt correctly during 3 minutes.

A.10 PROCESS FOR THE SPONTANEOUS DIVERGENT QA ASSESSMENT

To assess participants' spontaneous divergent QA, we give them an offline test. During this test, they are asked a short text, then write down all questions they have about it, within 2 minutes. We then investigate the quality of these questions as explained in Appendix B.2.

This test is administered two times, before and after the intervention for studies I (Chapter 3) and II (Chapter 4) where we only use "KidsAsk". For study III (Chapter 5), an additional measure is also collected after "KidsReflect".

The texts are the same for all participants but are changed for every measure (i.e. the texts are not the same for the initial, post-KidsReflect and post-KidsAsk measures).

A.11 PROCESS FOR THE THE METACOGNITIVE SENSITIVITY ASSESSMENT

To assess participants' metacognitive sensitivity, we give them an offline test. During this test, they are asked to answer a 12-item questionnaire and report their confidence levels in these answers. We then calculate a metacognitive sensitivity index using a 2 ROC function (as suggested in [63]).

This test is administered two times, before and after the intervention for studies I (Chapter 3) and II (Chapter 4) where we only use "KidsAsk". For study III (Chapter 5), an additional measure is also collected after "KidsReflect".

The items of the quizzes are the same for all participants but are changed for every measure (i.e. the items are not the same for the initial, post-KidsReflect and post-KidsAsk measures). See the tables below for the items and propositions for the three quizzes.

Test	Text
Initial	Insects have enormous eyes that go almost all the way round their heads. Thus, they can see forward, backward, right, left, up and down at the same time. Even better than a 3D vision headset! However, they don't see very clearly, because their eyes are made up of hundreds, or even thousands, of tiny single eyes. Each of these little eyes is responsible for processing a tiny part of the environment. Each little eye produces a single dot of a given hue and light intensity. These dots are then assembled together in the insect's brain to form a single, complete image. A bee's eye contains 4,500 tiny eyes. The eye of a dragonfly has 15,000: the most complex of all insect eyes!
Post-KidsReflect	Marie Curie was born in Poland. She studied at the Faculty of Science in Paris. She was one of the few women admitted. With her husband Pierre Curie, they work on the study of X-rays and discover two radioactive elements. The couple were awarded the Nobel Prize in Physics for this. During the First World War, Marie sets up ambulances equipped with X-ray machines, saving many wounded. Marie Curie was the first woman to win two Nobel Prizes: one with Pierre Curie in physics, and the other for her research in chemistry. She died in 1906.
Post-KidsAsk	Renewable energies are energies that can be renewed fairly quickly. They come from natural phenomena. They are called renewable or sustainable because they are the opposite of fossil fuels, which will eventually disappear as they are consumed in very large quantities and very quickly. Renewable energies are cleaner than fossil fuels, which are currently the most widely used energy sources worldwide. They are widely available around the globe and are "free". They are also less dangerous than fossil fuels.

Table 6: Texts used for the assessment of the spontaneous divergent QA behaviors during "KidsAsk" and "KidsReflect"

Question	Suggestion 1	Suggestion 2	Suggestion 3
Who invented the first vaccine?	Louis Pasteur	Albert Einstein	Gustave Eiffel
Who is the famous scientist who won two Nobel Prizes?	Rosalind Franklin	Marie Curie	Mileva Maric
Which of these objects is not a robot?	A drone	A stand-alone vacuum cleaner	A car
Can a robot learn things?	Yes, but only in one manner	No, it's just a machine	Yes, and in plenty manners
Where are most of the world's pandas found?	China	South Africa	Brazil
What is a T-Rex?	A baby dinosaur	A weak dinosaur	A super strong dinosaur
What is the sun?	A star	A planet	The biggest planet in the universe
How many planets are there in our solar system?	1	8	100
What makes it possible to create energy with water?	wind turbines	boat motors	windmills
Why are the northern ice bergs melting?	There is an earthquake	Polar bears make them melt	Pollution is raising temperatures
What are the Paralympics?	winter sports	sports for people with disabilities	sports that are not well known
What are e-sports?	video games competitions	internet competitions	video game programming competitions

Table 7: Quiz used to assess the initial measure of metacognitive sensitivity during the "KidsReflect" training

Question	Suggestion 1	Suggestion 2	Suggestion 3
Which scientist was the Tesla electric car named after?	NiKola Tesla	Albert Einstein	Louis Pasteur
Leonard de Vinci is	an Italian painter and scientist	a French musician	an American athlete
Which of these statements is false?	Robots only have a physical body	robots always need humans to be functional	robots have sensors
What is artificial intelligence?	working with a super-heavy robot	having machines with human-like intelligence	working with lots of different robots
Can animals talk to each other?	Yes, and in different ways	No	Yes, but only in one way
Where do we find the majority of elephants?	Europe	Africa	Northern America
What is the solar system?	A group pf stars	A group of planets, satellites, asteroids and comets	The Earth and the Sun together
What is an asteroid?	A star	A space body made of rocks and metals	A planet
What is global warming?	designates long-term variations in temperature	designates the disappearance of certain animals	designates the creation of new forests
What is the name of the action that consists in giving new life to waste?	Renovation	innovation	recycling
What is the Curling sport?	a sport practiced on ice	a sport practiced in a football stadium	a sport practiced in a swimming pool
How many strokes are there in swimming?	2	3	4

Table 8: Quiz used to assess the post-KidsReflect measure of metacognitive sensitivity"

Question	Suggestion 1	Suggestion 2	Suggestion 3
What is the profession of Albert Einstein?	Artist	Writer	Physicist
Alfred Nobel gives his name to the Nobel Prizes, the most prestigious scientific award. What did he invent?	The radio	a music instrument	dynamite
What is a computer algorithm?	A program given to a machine to tell it what to do	a sensor	a motor
What is a camera?	a visual sensor	a motor	a robot
Which animal has the best night sight?	cats	bats	boars
Where do white bears live?	Africa	Asia	Arctic
What is a comet?	A star	A planet	A body composed of ice and rock
What do scientists use to study the solar system?	A telescope	a camera	a microscope
Can we use wind as a source for energy?	yes	no	Yes, but it's a polluting energy source
Which of these energies is non-polluting?	Nuclear energy	coal	solar energy
What sport uses an oval ball?	Rugby	football	handball
Olympic games occur every	2 years	4 years	8 years

Table 9: Quiz used to assess the post-KidsAsk measure of metacognitive sensitivity"

B ANNOTATION GRIDS FOR ASSESSING CHILDREN'S AND LLMS' OUTPUT DURING THE INTERVENTIONS

During all of our intervention, we rely on manual annotations to assess the quality of participants' and GPT-3's performance, using standard grids inspired from literature.

It is to be noted that we used different scores to assess the divergence quality of the questions generated by children during our two studies: For the "KidsAsk" interventions in Part iii and Chapter 4, we use the percentage of divergent questions generated (we had a binary annotation: questions are either divergent or not). For the study reported in Chapter 5 where we combine "KidsReflect" and "KidsAsk", we use the average divergence level of the questions generated during the sessions; we had a 4-point continuous scale to determine this.

The reason for changing this measure was an attempt to have a more fine-grained measure of the divergent QA skills as having a binary classification was rather difficult for raters, in different situations.

B.1 CRITERIA FOR ACCEPTING A QUESTION DURING ALL STUDIES

For all trainings and tests proposed during this thesis, before analyzing the divergence dimension of a question entered by a participant, we first perform a manual assessment to see whether or not the question is 'acceptable'.

A question is considered acceptable if:

- It is a question — and not a statement.
- It is related to the educational resource in question.
- It is not repeated more than one time (repeated questions are only considered once).

A question is acceptable if the participant does not use the cues proposed by the agent they're interacting with. **Example:** For the text in Figure B.1, a linguistic cue '**What are the other**' and a semantic cue "**We can preserve food using fermentation**", we accepted questions such as:

- "What are the other possible ways to preserve food?" : used the agent propositions and is related to the text.

- “What are the other ways to protect our body?” : did not use the agent semantic proposition but is still related to the text.
- “How is the vaccine different from medication?” : did not use both of the agent propositions but is still related to the text.

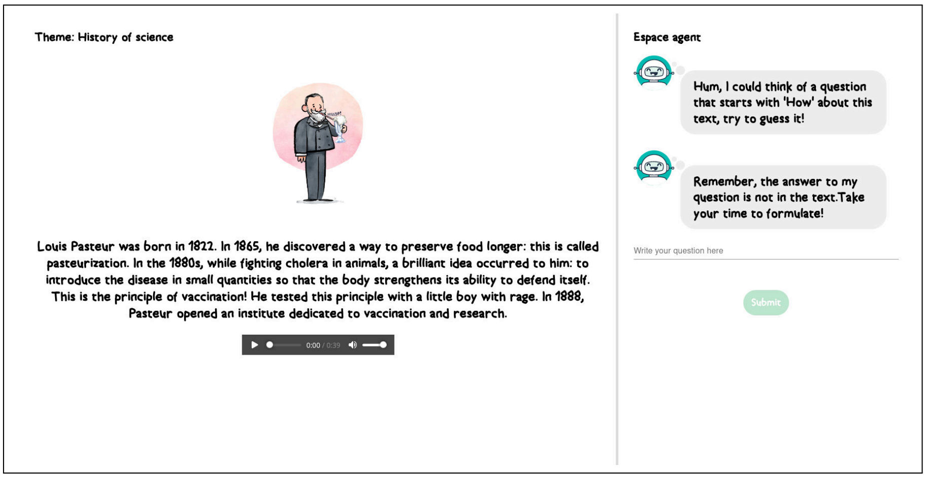


Figure B.1: Example of a text in "KidsAsk"

However, questions such as the following was not accepted:

- “It is another way to preserve food“ : this is a statement and not a question.
- "What are the other parts of a robot?" : this is not related to the Louis Pasteur text.
- "What are the other characters of the Simpsons?" : this is not a serious attempt.

The annotation of the divergence level only applies to acceptable questions.

B.2 ANNOTATING THE DIVERGENT QA PERFORMANCE IN STUDIES I AND II

For studies where we only use the "KidsAsk" platform, i.e. studies reported in Chapter 3 and Chapter 4, the divergent QA performance in a given work-space is computed as the percentage of divergent questions over all accepted questions generated by children in that work-space.

A question is considered divergent depending on whether or not its answer is already directly mentioned in the educational resource the participant is working on. **Example for the text in Figure B.1:**

- “What are the other possible ways to preserve food?” is considered divergent as its answer cannot be directly found in the text.

- "What is the aim of the Pasteurisation process?" is not considered divergent as the answer to it is explicitly stated in the text: "Longer preservation of food".

B.3 ANNOTATING THE DIVERGENT QA PERFORMANCE IN STUDY III

In study III where we combine "KidsReflect" with "KidsAsk", i.e. the study reported on in Chapter 5, we calculate the divergent QA performance as the average divergence score per question, over all questions generated during a given a given task, and judged acceptable.

To assign the divergence score per question, we use the following continuous grid:

- 0 if the answer to the question is explicitly stated in the text.
- 1 if the answer to the question is not explicitly stated in the text but can be implied from the text.
- 2 if he answer to the question is not explicitly stated in the text but can be guessed from the text.
- 3 if the answer to the question is not at all stated in the text.

We use the same process to assess the divergent QA performance during the "KidsAsk" training sessions and the offline tests we administer pre, post and in the middle of the "KidsReflect" intervention.

B.4 ANNOTATING THE SYNTACTIC QUALITY OF QUESTIONS DURING ALL STUDIES

In all studies where we perform a syntactic analysis of the questions generated by participants, we use a manual annotation process. The syntactic score of a question corresponds to the computed a question's score as the sum of the following criteria:

- One point if the question is high-level: the answer to this question is not a simple fact but requires to explain a mechanism, a relationship etc (example: 'Why were dinosaurs so big?'). 0 points if the question is low-level (example: 'How big is a dinosaur?').
- From 1 to 4 points, based on the syntactic construction of the question :
  1. 1 point for a 'closed' or declarative question (example: "Dinosaurs were big?").
  2. 2 points for questions with questioning words in the middle of the sentence (example: "The dinosaurs were how big?").

3. 3 points for a question without an interrogative formulation (example: "Why the dinosaurs are big ?").
4. 4 points for a questioning word in the beginning of the sentence that has interrogative syntax (example: "Why are dinosaurs big?").

- From 0 to 3 points, based on the use of questioning words :
  1. 0 point for a declarative question i.e., with no questioning word (example: "Dinosaurs were big?").
  2. 1 points for questions with 'Is/Are' (example: "Are dinosaurs big?").
  3. 2 points for the use of proper simple questioning words (example: "How big were the dinosaurs?").
  4. 3 points for the use of proper composite questioning words (example: "What difference ?").

Here, the maximum score for a question's quality is 8 points. In our studies, we report the average quality score over all questions generated by participants during a specific task, that were judged as 'acceptable'.

B.5 ANNOTATING THE PERFORMANCE DURING THE "KIDSREFLECT" PROCEDURAL TRAINING

The performance during the web-based app for the "KidsReflect" training was also assessed using a manual annotation process. This measures consists of calculating the percentage of *complete* and *correct* learning cycles that participants were able to achieve during this part of the training.

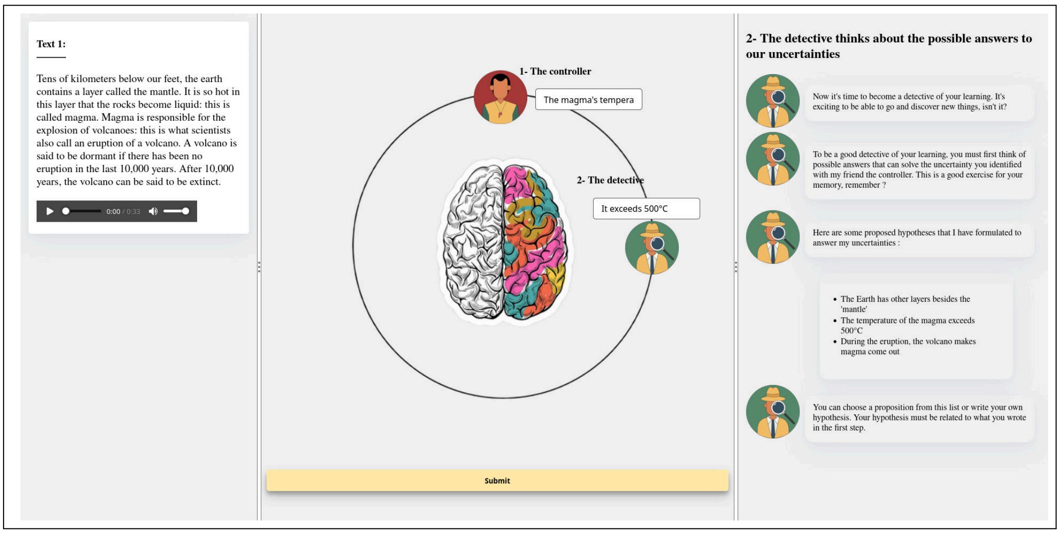


Figure B.2: Example of a task in part II of "KidsReflect"

A cycle is considered complete and correct if:

- All four steps (corresponding to the four curiosity-related metacognitive skills) are achieved.
- All four steps are coherent and related to each other semantically, i.e. they treat the same knowledge component.
- All four steps are considered acceptable by the annotators.

The four steps are considered acceptable following these criteria:

1. **The "IDENTIFY" skill with the first referee:** the entry is:
   - formulated as a declarative expression of an uncertainty.
   - contains a clear learning goal.
   - the expressed learning goal is not already mentioned in the text.
   - the expressed learning goal is semantically tightly related to the text.

   Example: For the text shown in Figure B.2, an entry such as "I don't know the temperature of the magma", "I don't know the other conditions for the eruption of a volcano" are accepted answers as they check all criteria above. However, entries such as "I don't know the the nature of the magma" or "I don't understand metacognition" are not accepted as they do not respect one or more of the conditions mentioned.

2. The "GUESS" skill with the detective: the entry is:
   - the guess is semantically tightly related to the text and knowledge gap identified in step 1.
   - the guess is not random, it is informed and formulated using former knowledge.

   Example: accepted answers are such as: "I know that water becomes liquid at 100 degrees so the Magma temperature should be higher" or "Maybe volcanoes require specific weather conditions to erupt". Examples of unaccepted answers are: "100 degrees", "volcanoes are very dangerous when they erupt", etc as they are not closer to random guesses than to informed ones.

3. The "SEEK" skill with the explorer: the entry is:
   - the question is semantically tightly related to the text and knowledge gap and guess generated in steps 1 and 2.
   - formulated as a question — and not a statement or declarative.
   - is divergent: its answer cannot be found directly in the text.

Example: accepted answers can be: "What is the temperature of the Magma?", "What other circumstances can lead to the eruption of a volcano?", etc. However, entries such as "Temperature for the Magma", "What is the Magma?", "What is metacognition?", etc are not accepted as they violate one or more of the conditions above.

4. The "ASSESS" skill with the second referee: the entry is:
    - the participant selects the correct answer to the question asked in step 3, in the case that this latter figures in the list of suggestions proposed by the agent.
    - the participant answers "I cannot find my answer" in the case that the answer does not figure in the list of suggestions proposed by the agent.

Once this process completed, we assess the number of correct and complete learning cycles and divide it by 8 (the number of tasks we proposed to children during our intervention). This gives us the percentage of success in this activity.

B.6 ANNOTATION GRIDS FOR ASSESSING GPT-3'S PERFORMANCE IN GENERATING PEDAGOGICAL CONTENT

During the study reported in Chapter 4, we use the LLM GPT-3 to generate the pedagogical content for our divergent QA training, i.e. the different cues proposed by the CA.

These cues consist of semantic and linguistic propositions. To help assess their quality, we perform manual annotations using specific procedures and grids as described below in order to compare their quality to what was hand-generated.

B.7 ANNOTATING THE GPT-3-DRIVEN LINGUISTIC CUES

In annotating the syntactic cues, i.e. the propositions of questioning words, we evaluate two dimensions: the variance of the cues proposed and their complexity level.

For the variance, we plot the histogram and assess the occurrence of the questioning words proposed. We then compare this to what was generated by hand.

For the complexity level, we assess this using a binary grid like the following:

- 0 points if the questioning word is simple, i.e. contains a single word, with no prepositions, etc. **Example:** what, where, who, etc.
- 1 point if the questioning word is compound, i.e. contains an association of words leading the question to be more precise and

high-level. **Example:** to what extent, what other, how different, etc.

We then compare the proportions of high-level questioning words within the cues proposed by the GPT-3-driven CA vs. the hand-crafted one.

B.8 ANNOTATING THE GPT-3-DRIVEN SEMANTIC CUES

In annotating the semantic cues, we evaluate two dimensions: their semantic relatedness to the educational resource and their divergence level with respect to it. For both dimensions, we had manual annotations processes following specific grids.

For the semantic relatedness, the score can go from 0 to 3, following this grid:

- 0 points if the cue does not at all relate to the text's general context, topic and theme. **Example for the text in Figure B.1:** "a robot can avoid obstacles using its sensors" for an incentive agent and "robots, sensors" for an open one.
- 1 point if the cue relates to the general topic of the text but not to its specific context. **Example for the text in Figure B.1:** "animals can have diseases like humans" for an incentive agent and "animals, sickness" for an open one.
- 2 points if the cue related to the specific context of the text but not to its key and main ideas. **Example for the text in Figure B.1:** "Louis Pasteur was born in France " for an incentive agent and "1822, birth" for an open one.
- 3 points if the cue is tightly related to a specific and key part of the text. **Example for the text in Figure B.1:** "freezing is another way of preserving food" for an incentive agent and "vaccines, medicine" for an open one.

Similarly, we had an annotation grid for the divergence level of the semantic cues. We only did this for the incentive cues (i.e. the short sentences that represent knowledge gaps) since the 'open' ones are separate words making the classification of the divergence level not relevant.

We used the following grid:

- 0 point if the cue is explicitly stated in the text. **Example for the text in Figure B.1:** "the process of pasteurization allows to preserve food longer."
- 1 points if the cue is not explicitly stated but can be implied from the text. **Example for the text in Figure B.1:** "the vaccination principle was a success".

- 2 points if the cue is not at all stated in the text. **Example for the text in Figure B.1:** "rage disease is caused by a virus".

C PEDAGOGICAL CONTENT FOR THE "KIDSREFLECT" DECLARATIVE TRAINING

C.1 SCRIPTS AND STORYBOARDS FOR THE VIDEOS

The scripts for the four videos we showed participants in the Part I of the "KidsReflect" training were written by the research team. They were then passed on to a motion designer to create the animated videos. Due to budget reasons, we only hired the designer for three videos. For the remaining video (video number 2 below), we used a powerpoint presentation with a voice over to create our own video.

The storyboards containing the scripts for the videos as well as the animation ideas used to realize them are available in Figures C.1, C.2, C.3 and C.4 below.

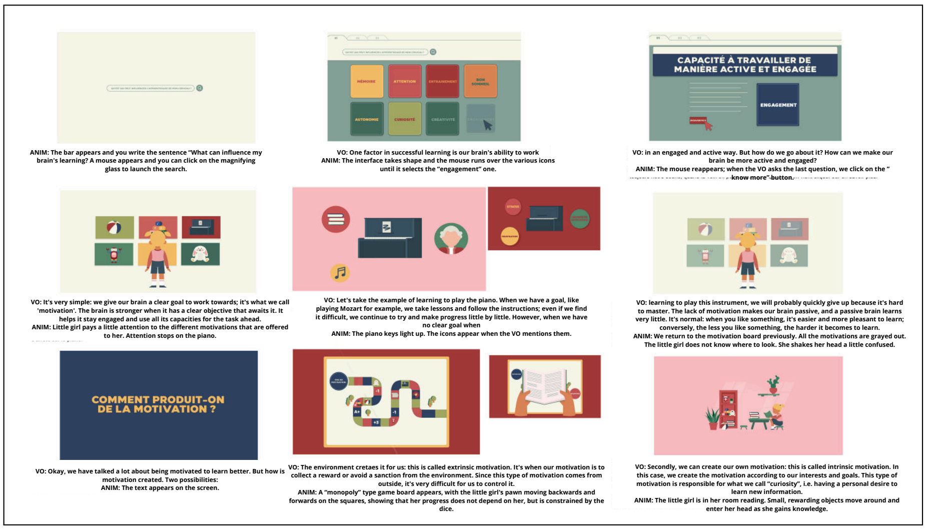


Figure C.1: Storyboard for video 1 in "KidsReflect": "Motivation types and curiosity"

Also, the videos —in French— are available for watching and downloading at https://www.dropbox.com/scl/fo/xycp0ew8qgx85ib4ddxq6/AIvkSk-mpAaix88E7dRq1y0?rlkey=ohrhbwbw2u9hutwq3f52e317g&st=pnvw5ao0&dl=0.

C.2 QUESTIONNAIRES FOR ASSESSING THE VIDEOS ACCESSIBILITY

To assess the comprehension of the videos, we had a 6-item questionnaire that we gave participants after the viewing session of each video. The questionnaires are available in the tables below.

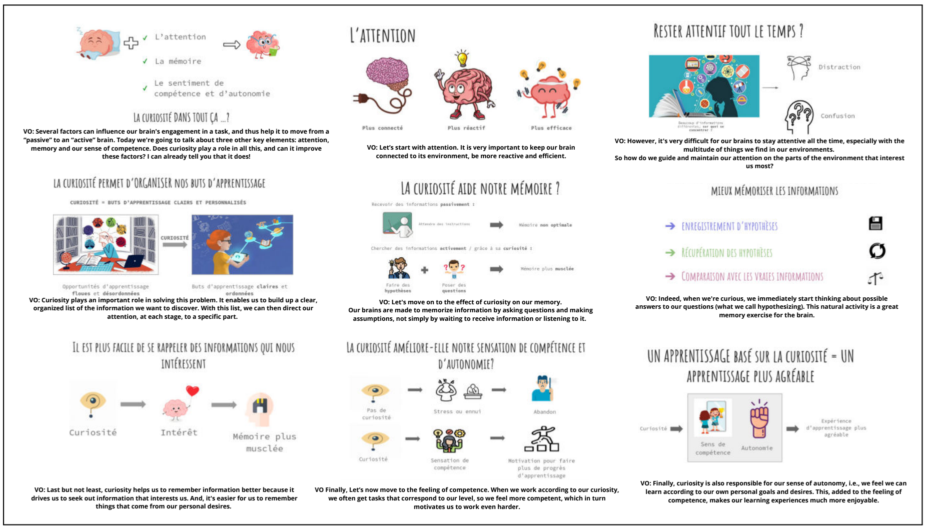


Figure C.2: Storyboard for video 2 in "KidsReflect": "Importance of curiosity during learning"

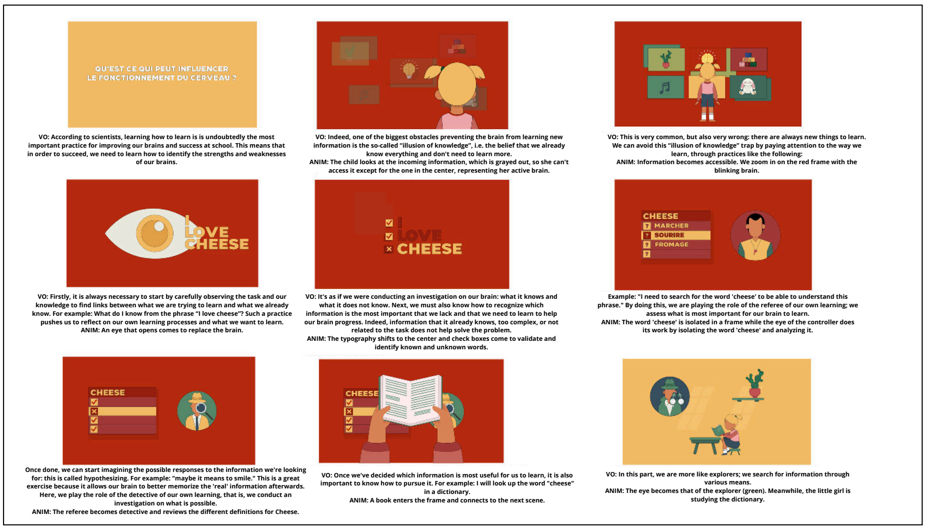


Figure C.3: Storyboard for video 3 in "KidsReflect": "Metacognition and basic skills"

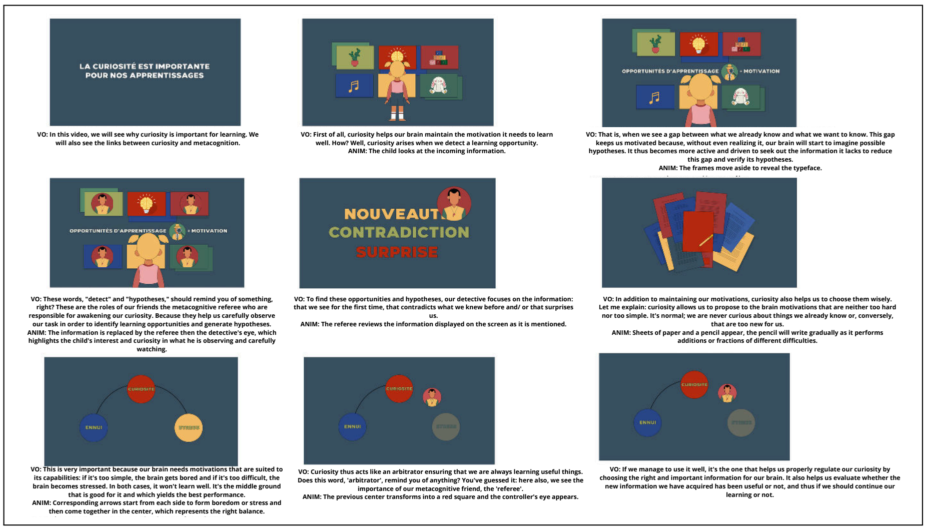


Figure C.4: Storyboard for video 4 in "KidsReflect": "Linking curiosity and metacognition"

Question	Suggestion 1	Suggestion 2	Suggestion 3
What is motivation?	Having a clear and precise goal that we want to pursue	It's the act of discovering new information	Changing our mind about what we want to learn
What does motivation give the brain?	The desire and strength to work	The ability to rest	New information
What happens when the brain lacks motivation?	The brain becomes active and learns better	Nothing happens	The brain becomes passive and learns very little
What is extrinsic motivation?	It consists in achieving personal satisfaction	It consists in collecting an external reward or avoiding punishment	It consists in having several motivations
What is intrinsic motivation?	It consists in achieving personal satisfaction	It consists in having someone else push us to learn something	It consists in collecting a reward or to avoid punishment
What is intrinsic motivation responsible for?	The ability to be calm	Stress during learning	Curiosity

Table 10: Questionnaire used to assess the comprehension of video 1 during part I of "KidsReflect": "Motivation types and curiosity"

Question	Suggestion 1	Suggestion 2	Suggestion 3
Attention allows the brain to be:	Less tired	Happier	More responsive, connected and efficient
Curiosity enables us to	direct our attention to the learning goals that interest us	think quicker	communicate better with others
Curiosity helps our memory because	Our brain memorizes better the information we are interested in and we look for on our own	We're less concentrated when we're curious	It's easier to remember difficult information
Curiosity helps us improve our sense of competence because :	It makes information easier	It helps us realize that we're making learning progress	It helps us learn easy information
When we follow our curiosity, we become :	More stressed in our learning	More autonomous in our learning	More fearful in our learning
Curiosity makes learning more enjoyable because :	We learn better when we feel autonomous and competent	It makes information easier	We have no choice about what we learn

Table 11: Questionnaire used to assess the comprehension of video 2 during part I of "KidsReflect": "Importance of curiosity during learning"

Question	Suggestion 1	Suggestion 2	Suggestion 3
What is the illusion of knowledge?	It's thinking we're not capable of learning new information	It's believing that our peers know more than we do	It's thinking that we already know everything and don't need to learn more
What is the first referee's role?	Decide what unknown information is important to learn	Memorize the information you already know	Ask new questions
What is the detective's role?	Memorize new information	Think of possible answers to the uncertainties you ask have	Choose what information you want to learn
What is the role of the explorer?	Evaluate the answers you find	Ask yourself what you want to learn	Ask the questions that lead to the answers resolve your uncertainties
What is the second referee's role?	Memorize new information	Evaluate whether You've found the information you need to answer your question	Ask what information you want to learn
What is metacogni-tion?	It's thinking about how we how we learn	It's learning as much new information as possible	It's watching how others learn new information

Table 12: Questionnaire used to assess the comprehension of video 3 during part I of "KidsReflect": "Metacognition and basic skills"

Question	Suggestion 1	Suggestion 2	Suggestion 3
What is a learning gap?	Information that is missing from our knowledge and that we want to learn	Information that we already know	Information that is stressing us out
How can the learning gap help our brain to become curious?	It generates stress in our brain	It generates boredom in our brain	It gives us motivation: we want to reduce this gap
What is responsible for awakening our curiosity?	Our parents or teachers	Information we already know	Our metacognitive referee, as it helps us see learning gaps
What regulates our curiosity?	Our parents or teachers	Our metacognitive referee, because it helps us choose the information that suits our brain	Stress and boredom
What kind of information do we choose to learn if we follow our curiosity?	Information that is very advanced in relation to our brain's capacities	Information that is adapted to our brain's capacities	Information that we already know
What keeps us curious?	Our metacognitive explorer, because it pushes us to discover new information	Stress or boredom	Not making learning progress

Table 13: Questionnaire used to assess the comprehension of video 4 during part I of "KidsReflect": "Linking curiosity and metacognition"

BIBLIOGRAPHY

[1] Rania Abdelghani, Edith Law, Chloé Desvaux, Pierre-Yves Oudeyer, and Hélène Sauzéon. "Interactive environments for training children's curiosity through the practice of metacognitive skills: a pilot study." In: *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 2023, pp. 495–501. URL: https://dl.acm.org/doi/abs/10.1145/3585088.3593880.

[2] Rania Abdelghani, Edith Law, Chloé Desvaux, Pierre-Yves Oudeyer, and Hélène Sauzéon. "Interactive environments for training children's curiosity through the practice of metacognitive skills: a pilot study." In: *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 2023, pp. 495–501.

[3] Rania Abdelghani, Pierre-Yves Oudeyer, Edith Law, Catherine de Vulpillières, and Hélène Sauzéon. "Conversational agents for fostering curiosity-driven learning in children." In: *International Journal of Human-Computer Studies* 167 (2022), p. 102887.

[4] Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. "Generative AI in the Classroom: Can Students Remain Active Learners?" In: *arXiv preprint* (2023). URL: https://arxiv.org/abs/2310.03192.

[5] Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. "Generative AI in the Classroom: Can Students Remain Active Learners?" In: *arXiv preprint* (2023).

[6] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. "Gpt-3-driven pedagogical agents to train children's curious question-asking skills." In: *International Journal of Artificial Intelligence in Education* (2023), pp. 1–36. URL: \url{https://link.springer.com/article/10.1007/s40593-023-00340-7}.

[7] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. "Gpt-3-driven pedagogical agents to train children's curious question-asking skills." In: *International Journal of Artificial Intelligence in Education* (2023), pp. 1–36.

[8] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. "Pedagogical agents for fostering question-asking skills in children." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–13.

[9] Vincent AWMM Aleven and Kenneth R Koedinger. "An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor." In: *Cognitive science* 26.2 (2002), pp. 147–179.

[10] Vincent Aleven, Kenneth R. Koedinger, and Karen Cross. "Tutoring Answer Explanation Fosters Learning with Understanding." In: *Artificial Intelligence in Education*. IOS Press, 1999, pp. 199–206.

[11] Vincent Aleven, Bruce McLaren, Ido Roll, and Kenneth Koedinger. "Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills." In: *Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Maceió, Alagoas, Brazil, August 30-September 3, 2004. Proceedings 7*. Springer. 2004, pp. 227–239.

[12] Dima Amso and Gaia Scerif. "The attentive brain: insights from developmental cognitive neuroscience." In: *Nature Reviews Neuroscience* 16.10 (2015), pp. 606–619.

[13] Roger Azevedo. "Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning." In: *Computers as Metacognitive Tools for Enhancing Learning*. Routledge, 2018, pp. 199–209.

[14] Roger Azevedo, Vincent Aleven, et al. *International handbook of metacognition and learning technologies*. Vol. 26. Springer, 2013.

[15] David Baidoo-Anu and Leticia Owusu Ansah. "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning." In: *Journal of AI* 7.1 (2023), pp. 52–62.

[16] Maria Bannert and Christoph Mengelkamp. "Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning?" In: *Metacognition and Learning* 3 (2008), pp. 39–58.

[17] Adrien Baranes, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. "Eye movements reveal epistemic curiosity in human observers." In: *Vision research* 117 (2015), pp. 81–90.

[18] Andrew G Barto. "Intrinsic motivation and reinforcement learning." In: *Intrinsically motivated learning in natural and artificial systems* (2013), pp. 17–47.

[19] Katarina Begus and Victoria Southgate. "Infant pointing serves an interrogative function." In: *Developmental science* 15.5 (2012), pp. 611–617.

[20] Katarina Begus and Victoria Southgate. "Curious learners: How infants' motivation to learn shapes and is shaped by infants' interactions with the social world." In: *Active learning from infancy to childhood: Social motivation, cognition, and linguistic mechanisms* (2018), pp. 13–37.

[21] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 610–623.

[22] Michael J Beran. *Foundations of metacognition*. Oxford University Press, 2012.

[23] Daniel E Berlyne. "Uncertainty and epistemic curiosity." In: *British Journal of Psychology* 53.1 (1962), pp. 27–34.

[24] Daniel E Berlyne. "Curiosity and Exploration: Animals spend much of their time seeking stimuli whose significance raises problems for psychology." In: *Science* 153.3731 (1966), pp. 25–33.

[25] Daniel Ellis Berlyne and Frances D Frommer. "Some determinants of the incidence and content of children's questions." In: *Child Development* (1966), pp. 177–189.

[26] Margherita Bernabei, Silvia Colabianchi, Andrea Falegnami, and Francesco Costantino. "Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances." In: *Computers and Education: Artificial Intelligence* 5 (2023), p. 100172.

[27] Som Biswas. "Role of Chat GPT in Education." In: *Available at SSRN 4369981* (2023).

[28] Robert A Bjork, John Dunlosky, and Nate Kornell. "Self-regulated learning: Beliefs, techniques, and illusions." In: *Annual review of psychology* 64 (2013), pp. 417–444.

[29] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. "Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 1004–1015.

[30] Inge Bretherton and Marjorie Beeghly. "Talking about internal states: The acquisition of an explicit theory of mind." In: *Developmental psychology* 18.6 (1982), p. 906.

[31] Ingar Brinck and Rikard Liljenfors. "The developmental origin of metacognition." In: *Infant and Child Development* 22.1 (2013), pp. 85–101.

[32] Garvin Brod and Jasmin Breitwieser. "Lighting the wick in the candle of learning: generating a prediction stimulates curiosity." In: *NPJ science of learning* 4.1 (2019), p. 17.

[33] Jonathan Brown, Gwen A. Frishkoff, and Maxine Eskénazi. "Automatic Question Generation for Vocabulary Assessment." In: *HLT*. 2005.

[34] Tom B. Brown et al. "Language Models are Few-Shot Learners." In: *ArXiv* abs/2005.14165 (2020).

[35] Maja Brydevall, Daniel Bennett, Carsten Murawski, and Stefan Bode. "The neural encoding of information prediction errors during non-instrumental information seeking." In: *Scientific reports* 8.1 (2018), pp. 1–11.

[36] Janis A Bulgren, Janet G Marquis, B Keith Lenz, Donald D Deshler, and Jean B Schumaker. "The effectiveness of a question-exploration routine for enhancing the content learning of secondary students." In: *Journal of Educational Psychology* 103.3 (2011), p. 578.

[37] Jill Burstein. "Opportunities for Natural Language Processing Research in Education." In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 6–27. ISBN: 978-3-642-00382-0.

[38] Aimee A Callender, Ana M Franco-Watkins, and Andrew S Roberts. "Improving metacognition in the classroom through instruction, training, and feedback." In: *Metacognition and learning* 11 (2016), pp. 215–235.

[39] Peter Carruthers. "Basic questions." In: *Mind & Language* 33.2 (2018), pp. 130–147.

[40] Jessy Ceha, Nalin Chhibber, Joslin Goh, Corina McDonald, Pierre-Yves Oudeyer, Dana Kulić, and Edith Law. "Expression of curiosity in social robots: Design, perception, and effects on behaviour." In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–12.

[41] Canyu Chen and Kai Shu. "Can LLM-Generated Misinformation Be Detected?" In: *arXiv preprint arXiv:2309.13788* (2023).

[42] M. M Chouinard. "Children's questions: A mechanism for cognitive development: III. Diary study of children's questions." In: *Monographs of the Society for Research in Child Development* 72 (2004), pp. 45–57. DOI: https://doi.org/10.1111/j.1540-5834.2007.00415.x.

[43] Michelle M Chouinard, Paul L Harris, and Michael P Maratsos. "Children's questions: A mechanism for cognitive development." In: *Monographs of the society for research in child development* (2007), pp. i–129.

[44] Benjamin Clement, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes. "Multi-armed bandits for intelligent tutoring systems." In: *arXiv preprint arXiv:1310.3174* (2013).

[45] Benjamin Clément, Hélène Sauzéon, Didier Roy, and Pierre-Yves Oudeyer. "Improved Performances and Motivation in Intelligent Tutoring Systems: Combining Machine Learning and Learner Choice." In: *arXiv preprint arXiv:2402.01669* (2024).

[46] Benjamin Clément. "Adaptive Personalization of Pedagogical Sequences using Machine Learning." PhD thesis. University of Bordeaux, 2018.

[47] Diana I Cordova and Mark R Lepper. "Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice." In: *Journal of educational psychology* 88.4 (1996), p. 715.

[48] Ali Darvishi, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. "Impact of AI assistance on student agency." In: *Computers & Education* 210 (2024), p. 104967.

[49] Peter Dayan and Yael Niv. "Reinforcement learning: the good, the bad and the ugly." In: *Current opinion in neurobiology* 18.2 (2008), pp. 185–196.

[50] Edward L Deci and Richard M Ryan. "Curiosity and Self-Directed Learning: The Role of Motivation in Education." In: (1981).

[51] Edward L Deci and Richard M Ryan. "The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior." In: *Psychological inquiry* 11.4 (2000), pp. 227–268.

[52] Adolph J Delgado, Liane Wardlow, Katherine McKnight, and Kimberly O'Malley. "Educational technology: A review of the integration, resources, and effectiveness of technology in k-12 classrooms." In: *Journal of Information Technology Education* 14 (2015).

[53] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[54] Adam Kenneth Dubé and Run Wen. "Identification and evaluation of technology trends in K-12 education from 2011 to 2021." In: *Education and information technologies* 27.2 (2022), pp. 1929–1958.

[55] Rachit Dubey and Thomas L Griffiths. "Reconciling novelty and complexity through a rational analysis of curiosity." In: *Psychological Review* 127.3 (2020), p. 455.

[56] John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. "Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology." In: *Psychological Science in the Public interest* 14.1 (2013), pp. 4–58.

[57] Carol S Dweck. "Mind-sets." In: *Principal leadership* 10.5 (2010), pp. 26–29.

[58] Susan Engel. "Children's need to know: Curiosity in schools." In: *Harvard educational review* 81.4 (2011), pp. 625–645.

[59] Yana Fandakova and Matthias J Gruber. "States of curiosity and interest enhance memory differently in adolescents and in children." In: *Developmental Science* 24.1 (2021), e13005.

[60] Lily FitzGibbon, Johnny King L Lau, and Kou Murayama. "The seductive lure of curiosity: Information as a motivationally salient reward." In: *Current Opinion in Behavioral Sciences* 35 (2020), pp. 21–27.

[61] John H Flavell. "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry." In: *American psychologist* 34.10 (1979), p. 906.

[62] John H Flavell, Frances L Green, Eleanor R Flavell, Paul L Harris, and Janet Wilde Astington. "Young children's knowledge about thinking." In: *Monographs of the society for research in child development* (1995), pp. i–113.

[63] Stephen M Fleming and Hakwan C Lau. "How to measure metacognition." In: *Frontiers in human neuroscience* 8 (2014), p. 443.

[64] Sébastien Forestier, Yoan Mollard, and Pierre-Yves Oudeyer. "Intrinsically Motivated Goal Exploration Processes with Automatic Curriculum Learning.(aug 2017)." In: *arXiv preprint arXiv:1708.02190* (2017).

[65] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. "Active learning increases student performance in science, engineering, and mathematics." In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8410–8415.

[66] Thomas L Friedman. *The world is flat: A brief history of the twenty-first century*. Macmillan, 2005.

[67] Thomas B Frøsig. "Expanding the Technology Acceptance Model (TAM) to Consider Teachers Needs and Concerns in the Design of Educational Technology (EdTAM)." In: *International Journal of Emerging Technologies in Learning* 18.16 (2023).

[68] Pierre Paul Gagné and Louis-Philippe Longpré. *Apprendre... avec Réflecto*. Chenelière Éducation, 2004.

[69] J. J. Gallagher and M. J. Aschner. "A preliminary report on analyses of classroom interaction." In: *Merrill-Palmer Quarterly* 9 (1963), pp. 183–194.

[70] Susan J Galvin, John V Podd, Vit Drga, and John Whitmore. "Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions." In: *Psychonomic bulletin & review* 10 (2003), pp. 843–876.

[71] Russell Golman and George Loewenstein. "Information gaps: A theory of preferences regarding the presence and absence of information." In: *Decision* 5.3 (2018), p. 143.

[72] Alison Gopnik and Andrew N Meltzoff. *Words, thoughts, and theories*. Mit Press, 1997.

[73] Goren Gordon, Cynthia Breazeal, and Susan Engel. "Can Children Catch Curiosity from a Social Robot?" In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: Association for Computing Machinery, 2015, 91–98. DOI: 10.1145/2696454.2696469. URL: https://doi.org/10.1145/2696454.2696469.

[74] Adele Eskeles Gottfried, Kathleen Suzanne Johnson Preston, Allen W Gottfried, Pamella H Oliver, Danielle E Delany, and Sirena M Ibrahim. "Pathways from parental stimulation of children's curiosity to high school science course accomplishments and science career interest and skill." In: *International Journal of Science Education* 38.12 (2016), pp. 1972–1995.

[75] Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. "Information-seeking, curiosity, and attention: computational and neural mechanisms." In: *Trends in cognitive sciences* 17.11 (2013), pp. 585–593.

[76] Louise Goupil and Sid Kouider. "Developing a reflective mind: From core metacognition to explicit self-reflection." In: *Current Directions in Psychological Science* 28.4 (2019), pp. 403–408.

[77] Louise Goupil and Joëlle Proust. "Curiosity as a metacognitive feeling." In: *Cognition* 231 (2023), p. 105325.

[78] Louise Goupil, Margaux Romand-Monnier, and Sid Kouider. "Infants ask for help when they know they don't know." In: *Proceedings of the National Academy of Sciences* 113.13 (2016), pp. 3492–3496.

[79] Arthur C Graesser and Cathy L McMahen. "Anomalous information triggers questions when adults solve quantitative problems and comprehend stories." In: *Journal of Educational Psychology* 85.1 (1993), p. 136.

[80] Arthur C. Graesser and Natalie K Person. "Question Asking During Tutoring." In: *American Educational Research Journal* 31 (1994), pp. 104–137. DOI: 10.3102/00028312022002217.

[81] Wendy S Grolnick and Richard M Ryan. "Autonomy support in education: Creating the facilitating environment." In: *New directions in educational psychology*. Routledge, 2013, pp. 213–231.

[82] Madeleine E Gross, Claire M Zedelius, and Jonathan W Schooler. "Cultivating an understanding of curiosity as a seed for creativity." In: *Current Opinion in Behavioral Sciences* 35 (2020), pp. 77–82.

[83] Matthias J Gruber and Yana Fandakova. "Curiosity in childhood and adolescence—what can we learn from the brain." In: *Current Opinion in Behavioral Sciences* 39 (2021), pp. 178–184.

[84] Matthias J Gruber, Bernard D Gelman, and Charan Ranganath. "States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit." In: *Neuron* 84.2 (2014), pp. 486–496.

[85] Matthias J Gruber and Charan Ranganath. "How curiosity enhances hippocampus-dependent memory: The prediction, appraisal, curiosity, and exploration (PACE) framework." In: *Trends in cognitive sciences* 23.12 (2019), pp. 1014–1025.

[86] Frédéric Guay. "Applying Self-Determination Theory to Education: Regulations Types, Psychological Needs, and Autonomy Supporting Behaviors." In: *Canadian Journal of School Psychology* 37.1 (2022), pp. 75–92. DOI: 10.1177/08295735211055355. eprint: https://doi.org/10.1177/08295735211055355. URL: https://doi.org/10.1177/08295735211055355.

[87] Sabrina Habib, Thomas Vogel, Xiao Anli, and Evelyn Thorne. "How does generative artificial intelligence impact student creativity?" In: *Journal of Creativity* 34.1 (2024), p. 100072.

[88] Douglas J Hacker, John Dunlosky, and Arthur C Graesser. *Handbook of metacognition in education*. Routledge, 2009.

[89] Abid Haleem, Mohd Javaid, Mohd Asim Qadri, and Rajiv Suman. "Understanding the role of digital technologies in education: A review." In: *Sustainable Operations and Computers* 3 (2022), pp. 275–285.

[90] Harry F Harlow, Margaret Kuenne Harlow, and Donald R Meyer. "Learning motivated by a manipulation drive." In: *Journal of experimental psychology* 40.2 (1950), p. 228.

[91] Paul L Harris, Deborah T Bartz, and Meredith L Rowe. "Young children communicate their ignorance and ask questions." In: *Proceedings of the National Academy of Sciences* 114.30 (2017), pp. 7884–7891.

[92] S. G. Hart and L. E. Staveland. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." In: 1988, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[93] Ted S Hasselbring. "Research on the effectiveness of computer-based instruction: A review." In: *International Review of Education* 32 (1986), pp. 313–324.

[94] Fritz Heider and Marianne Simmel. "An experimental study of apparent behavior." In: *The American journal of psychology* 57.2 (1944), pp. 243–259.

[95] Bruce Henderson and Shirley G Moore. "Children's responses to objects differing in novelty in relation to level of curiosity and adult behavior." In: *Child development* (1980), pp. 457–465.

[96] Lisa M Henjens. "The use of think-aloud strategies to solve word problems." MA thesis. University of Nebraska - Lincoln, Lincoln, NE, 2007.

[97] Suzanne Hidi and K Ann Renninger. "The four-phase model of interest development." In: *Educational psychologist* 41.2 (2006), pp. 111–127.

[98] Steven Higgins, ZhiMin Xiao, and Maria Katsipataki. "The Impact of Digital Technology on Learning: A Summary for the Education Endowment Foundation. Full Report." In: *Education Endowment Foundation* (2012).

[99] Philip A Higham, Timothy J Perfect, and Davide Bruno. "Investigating strength and frequency effects in recognition memory using type-2 signal detection theory." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.1 (2009), p. 57.

[100] Brian Hill. "Taking the help or going alone: ChatGPT and class assignments." In: *HEC Paris Research Paper Forthcoming* (2023).

[101] Robert Aubrey Hinde. "Factors governing the changes in strength of a partially inborn response, as shown by the mobbing behaviour of the chaffinch (Fringilla coelebs)-I. The nature of the response, and an examination of its course." In: *Proceedings of the Royal Society of London. Series B-Biological Sciences* 142.908 (1954), pp. 306–331.

[102] Paul A Howard-Jones and Skevi Demetriou. "Uncertainty and engagement with learning games." In: *Instructional Science* 37 (2009), pp. 519–536.

[103] Jean Humphries and Molly Ness. "Beyond who, what, where, when, why, and how: Preparing students to generate questions in the age of common core standards." In: *Journal of Research in Childhood Education* 29.4 (2015), pp. 551–564.

[104] William James. "On a certain blindness in human beings." In: (1899).

[105] William James. *Talks to Teachers on Psychology and to Students on Some of Life's Ideals*. Vol. 12. Harvard University Press, 1983.

[106] William James. *The principles of psychology*. Vol. 1. Cosimo, Inc., 2007.

[107] Lan Jiang, Chunliang Yang, Zhongling Pi, Yangping Li, Shaohang Liu, and Xinfa Yi. "Individuals with High Metacognitive Ability Are Better at Divergent and Convergent Thinking." In: *Journal of Intelligence* 11.8 (2023), p. 162.

[108] Jamie J Jirout. "Supporting early scientific thinking through curiosity." In: *Frontiers in Psychology* 11 (2020), p. 539140.

[109] Jamie J Jirout, Virginia E Vitiello, and Sharon K Zumbrunn. "Curiosity in schools." In: *The new science of curiosity* 1.1 (2018), pp. 243–266.

[110] Jamie Jirout and David Klahr. "Children's Question Asking and Curiosity: A Training Study Conference." In: *Proceedings of the Society for Research on Educational Effectiveness Conference*. 2011, pp. 1–4.

[111] Jamie Jirout and David Klahr. "Children's scientific curiosity: In search of an operational definition of an elusive concept." In: *Developmental review* 32.2 (2012), pp. 125–160.

[112] Aidan Jones, Susan Bull, and Ginevra Castellano. ""I Know That Now, I'm Going to Learn This Next" Promoting Self-regulated Learning with a Robotic Tutor." In: *International Journal of Social Robotics* 10 (2018), pp. 439–454. DOI: 10.1007/s12369-017-0430-y.

[113] Angela Jones, Nora Swaboda, and Azzurra Ruggeri. "Developmental changes in question-asking." In: *The questioning child: Insights from psychology and education* (2020), pp. 118–143.

[114] Brett D Jones. "Motivating students to engage in learning: the MUSIC model of academic motivation." In: *International Journal of Teaching and Learning in Higher Education* 21.2 (2009), pp. 272–285.

[115] John T. Jost, Arie W. Kruglanski, and Thomas O. Nelson. "Social Metacognition: An Expansionist Review." In: *Personality and Social Psychology Review* 2.2 (1998), pp. 137–154. DOI: 10.1207/s15327957pspr0202\textunderscore6.

[116] Reyes Juárez-Ramírez, Raúl Navarro-Almanza, Yail Gomez-Tagle, Guillermo Licea, Carlos Huertas, and German Quinto. "Orchestrating an adaptive intelligent tutoring system: towards integrating the user profile for learning improvement." In: *Procedia-Social and Behavioral Sciences* 106 (2013), pp. 1986–1999.

[117] Min Jeong Kang, Ming Hsu, Ian M Krajbich, George Loewenstein, Samuel M McClure, Joseph Tao-yi Wang, and Colin F Camerer. "The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory." In: *Psychological science* 20.8 (2009), pp. 963–973.

[118] Todd B Kashdan, Fallon R Goodman, David J Disabato, Patrick E McKnight, Kerry Kelso, and Carl Naughton. "Curiosity has comprehensive benefits in the workplace: Developing and validating a multidimensional workplace curiosity scale in United States and German employees." In: *Personality and Individual Differences* 155 (2020), p. 109717.

[119] Todd B Kashdan and John E Roberts. "Social anxiety's impact on affect, curiosity, and social self-efficacy during a high self-focus social threat situation." In: *Cognitive Therapy and Research* 28 (2004), pp. 119–141.

[120] Todd B Kashdan and Paul J Silvia. "Curiosity and interest: The benefits of thriving on novelty and challenge." In: *Oxford handbook of positive psychology* 2 (2009), pp. 367–374.

[121] Todd B Kashdan, Melissa C Stiksma, David J Disabato, Patrick E McKnight, John Bekier, Joel Kaji, and Rachel Lazarus. "The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people." In: *Journal of Research in Personality* 73 (2018), pp. 130–149.

[122] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." In: *Learning and individual differences* 103 (2023), p. 102274.

[123] Greg P Kearsley. "Questions and question asking in verbal discourse: A cross-disciplinary review." In: *Journal of psycholinguistic research* 5 (1976), pp. 355–375.

[124] Celeste Kidd and Abeba Birhane. "How AI can distort human beliefs." In: *Science* 380.6651 (2023), pp. 1222–1223.

[125] Celeste Kidd and Benjamin Y Hayden. "The psychology and neuroscience of curiosity." In: *Neuron* 88.3 (2015), pp. 449–460.

[126] Celeste Kidd, Steven T Piantadosi, and Richard N Aslin. "The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex." In: *PloS one* 7.5 (2012), e36399.

[127] Royce Kimmons, Joshua Rosenberg, and Bohdana Allman. "Trends in educational technology: What Facebook, Twitter, and Scopus can tell us about current research and practice." In: *TechTrends* 65 (2021), pp. 125–136.

[128] Asher Koriat, Ravit Nussinson, Herbert Bless, and Nira Shaked. "Information-based and experience-based metacognitive judgments: Evidence from subjective confidence." In: *Handbook of metamemory and memory* (2008), pp. 117–135.

[129] Nate Kornell and Robert A. Bjork. "Optimising self-regulated study: The benefits—and costs—of dropping flashcards." In: *Journal ofExperimental Psychology: Learning, Memory, and Cognition* 42 (2016), pp. 914–924.

[130] David R Krathwohl. "A revision of Bloom's taxonomy: An overview." In: *Theory into practice* 41.4 (2002), pp. 212–218.

[131] Chen-Lin C Kulik and James A Kulik. "Effectiveness of computer-based instruction: An updated analysis." In: *Computers in human behavior* 7.1-2 (1991), pp. 75–94.

[132] Sawan Kumar and Partha P. Talukdar. "Reordering Examples Helps during Priming-based Few-Shot Learning." In: *CoRR* abs/2106.01751 (2021). arXiv: 2106.01751. URL: https://arxiv.org/abs/2106.01751.

[133] Tianran LI and Guoliang YU. "Why are humans curious? Concept, functions and theories of interpersonal curiosity." In: *Advances in Psychological Science* 23.1 (2015), p. 132.

[134] Emily R Lai. "Metacognition: A literature review." In: *Always learning: Pearson research report* 24 (2011), pp. 1–40.

[135] Johnny King L Lau, Hiroki Ozono, Kei Kuratomi, Asuka Komiya, and Kou Murayama. "Hunger for knowledge: how the irresistible lure of curiosity is generated in the brain." In: *bioRxiv* (2018), p. 473975.

[136] Marco Lauriola, Jordan A Litman, Patrick Mussel, Riccardo De Santis, Howard M Crowson, and Robert R Hoffman. "Epistemic curiosity and self-regulation." In: *Personality and Individual Differences* 83 (2015), pp. 202–207.

[137] Rebecca Lawson. "The science of cycology: Failures to understand how everyday objects work." In: *Memory & Cognition* 34.8 (2006), pp. 1667–1675. DOI: 10.3758/BF03195929. URL: https://doi.org/10.3758/BF03195929.

[138] Pierre Lefavrais. *Alouette-R : [test d'analyse de la vitesse en lecture a partir d'un texte]*. Paris : Centre de psychologie appliquee, 2005.

[139] Angeline Lillard and Nicole Else-Quest. "Evaluating montessori education." In: *science* 313.5795 (2006), pp. 1893–1894.

[140] Diane Litman. "Natural Language Processing for Enhancing Teaching and Learning." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (2016). DOI: 10.1609/aaai.v30i1.9879. URL: https://ojs.aaai.org/index.php/AAAI/article/view/9879.

[141] Jordan A Litman. "Interest and deprivation factors of epistemic curiosity." In: *Personality and individual differences* 44.7 (2008), pp. 1585–1595.

[142] Jordan A Litman. "Curiosity and metacognition." In: *Metacognition: New research developments* 105 (2009), p. 116.

[143] Jordan A Litman and Tiffany L Jimerson. "The measurement of curiosity as a feeling of deprivation." In: *Journal of personality assessment* 82.2 (2004), pp. 147–157.

[144] Jordan A. Litman and Charles D. Spielberger. "Measuring Epistemic Curiosity and Its Diversive and Specific Components." In: *Journal of Personality Assessment* 80.1 (2003). PMID: 12584070, pp. 75–86. DOI: 10.1207/S15327752JPA8001\textunderscore16.

[145] Jordan Litman. "Curiosity and the pleasures of learning: Wanting and liking new information." In: *Cognition & emotion* 19.6 (2005), pp. 793–814.

[146] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." In: *arXiv preprint arXiv:2107.13586* (2021).

[147] George Loewenstein. "The psychology of curiosity: A review and reinterpretation." In: *Psychological bulletin* 116.1 (1994), p. 75.

[148] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL: https://aclanthology.org/2022.acl-long.556.

[149] Kelsey Lucca and Makeba Parramore Wilbourn. "The what and the how: Information-seeking pointing gestures facilitate learning labels and functions." In: *Journal of Experimental Child Psychology* 178 (2019), pp. 417–436.

[150] David M Lydon-Staley, Dale Zhou, Ann Sizemore Blevins, Perry Zurn, and Danielle S Bassett. "Hunters, busybodies and the knowledge network building associated with deprivation curiosity." In: *Nature human behaviour* 5.3 (2021), pp. 327–336.

[151] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. *MathDial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems*. 2023. arXiv: 2305.14536 [cs.CL].

[152] Mohsen Mahdavi. "An overview: Metacognition in education." In: *International Journal of Multidisciplinary and current research* 2.6 (2014), pp. 529–535.

[153] Mona H Mahmoud and Sanaa H Abo El-Hamayed. "An intelligent tutoring system for teaching the grammar of the Arabic language." In: *Journal of Electrical Systems and Information Technology* 3.2 (2016), pp. 282–294.

[154] Brian Maniscalco and Hakwan Lau. "A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings." In: *Consciousness and cognition* 21.1 (2012), pp. 422–430.

[155] van Loon Mariëtte and Janneke van de Pol. "Judging own and peer performance when using feedback in elementary school." In: *Learning and Individual Differences* Volume 74 (2019), pp. 101–754. URL: https://doi.org/10.1016/j.lindif.2019.101754.

[156] Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. "GPTeach: Interactive TA Training with GPT Based Students." In: (2023).

[157] Andrew J Martin. "Courage in the classroom: Exploring a new framework predicting academic performance and engagement." In: *School Psychology Quarterly* 26.2 (2011), p. 145.

[158] Alexandra Marx, Urs Fuhrer, and Terry Hartig. "Effects of classroom seating arrangements on children's question-asking." In: *Learning Environments Research* 2 (1999), pp. 249–263.

[159] Janet Metcalfe, Bennett L Schwartz, and Teal S Eich. "Epistemic curiosity and the region of proximal learning." In: *Current Opinion in Behavioral Sciences* 35 (2020). Curiosity (Explore vs Exploit), pp. 40–47. ISSN: 2352-1546. DOI: https://doi.org/10.1016/j.cobeha.2020.06.007. URL: https://www.sciencedirect.com/science/article/pii/S2352154620300954.

[160] Tyler M Miller and Lisa Geraci. "Training metacognition in the classroom: The influence of incentives and feedback on exam predictions." In: *Metacognition and Learning* 6 (2011), pp. 303–314.

[161] Candice M Mills, Cristine H Legare, Meridith G Grant, and Asheley R Landrum. "Determining who to question, what to ask, and how much information to ask for: The development of inquiry in young children." In: *Journal of Experimental Child Psychology* 110.4 (2011), pp. 539–560.

[162] Candice M Mills, Kaitlin R Sands, Sydney P Rowles, and Ian L Campbell. ""I want to know more!": Children are sensitive to explanation quality when exploring new information." In: *Cognitive Science* 43.1 (2019), e12706.

[163] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" In: *arXiv preprint arXiv:2202.12837* (2022).

[164] Gaia Molinaro, Irene Cogliati Dezza, Sarah Katharina Bühler, Christina Moutsiana, and Tali Sharot. "Multifaceted information-seeking motives in children." In: *Nature Communications* 14.1 (2023), p. 5505.

[165] Maria Montessori. *The montessori method*. Transaction publishers, 2013.

[166] Steven Moore, Huy A Nguyen, Tianying Chen, and John Stamper. "Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods." In: *European Conference on Technology Enhanced Learning*. Springer. 2023, pp. 229–245.

[167] K. Murayama, A. B. Blake, T. Kerr, and A. D. Castel. "When enough is not enough: information overload and metacognitive decisions to stop studying information." In: *Memory* 16:2 (2008), pp. 125–136. DOI: 10.1080/09658210701763899.

[168] Kou Murayama, Lily FitzGibbon, and Michiko Sakaki. "Process account of curiosity and interest: A reward-learning perspective." In: *Educational Psychology Review* 31 (2019), pp. 875–895.

[169] Debra Myhill and Frances Dunkin. "Questioning learning." In: *Language and education* 19.5 (2005), pp. 415–427.

[170] Thomas O Nelson. "Consciousness and metacognition." In: *American psychologist* 51.2 (1996), p. 102.

[171] Lynn Newton. *Questioning: a window on productive thinking*. International Centre for Innovation in Education, 2017.

[172] Ageliki Nicolopoulou. "Play, cognitive development, and the social world: Piaget, Vygotsky, and beyond." In: *Human development* 36.1 (1993), pp. 1–23.

[173] Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. *The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters' Exam Performances*. Tech. rep. Center for Open Science, 2024.

[174] Evan Orticio, Martin Meyer, and Celeste Kidd. "Children flexibly adapt their evidentiary standards to their informational environment." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 45. 45. 2023.

[175] P-Y Oudeyer, Jacqueline Gottlieb, and Manuel Lopes. "Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies." In: *Progress in brain research* 229 (2016), pp. 257–284.

[176] Pierre-Yves Oudeyer and Frederic Kaplan. "What is intrinsic motivation? A typology of computational approaches." In: *Frontiers in neurorobotics* 1 (2007), p. 6.

[177] Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner. "Intrinsic Motivation Systems for Autonomous Mental Development." In: *IEEE Transactions on Evolutionary Computation* 11.2 (2007), pp. 265–286. DOI: 10.1109/TEVC.2006.890271.

[178] Nesrin Ozturk. "Assessing metacognition: Theory and practices." In: *International Journal of Assessment Tools in Education* 4.2 (2017), pp. 134–148.

[179] Emily Grossnickle Peterson. "Supporting curiosity in schools and classrooms." In: *Current Opinion in Behavioral Sciences* 35 (2020), pp. 7–13.

[180] Jean Piaget. *La naissance de l'intelligence chez l'enfant*. Delachaux et Niestlé Neuchatel-Paris, 1935.

[181] Paul R Pintrich et al. "A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)." In: (1991).

[182] F Poli, G Serino, RB Mars, and S Hunnius. "Infants tailor their attention to maximize learning." In: *Science advances* 6.39 (2020), eabb5053.

[183] Tim Post and Juliette H Walma van der Molen. "Development and validation of a questionnaire to measure primary school children's images of and attitudes towards curiosity (the CIAC questionnaire)." In: *Motivation and emotion* 43 (2019), pp. 159–178.

[184] Rosalind Potts, Gabriella Davies, and David R Shanks. "The benefit of generating errors during learning: What is the locus of the effect?" In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45.6 (2019), p. 1023.

[185] Joëlle Proust. "Metacognition and mindreading: one or two functions." In: *Foundations of metacognition* (2012), pp. 234–251.

[186] Taffy E Raphael and P David Pearson. "Increasing students' awareness of sources of information for answering questions." In: *American Educational Research Journal* 22.2 (1985), pp. 217–235.

[187] Chris Reinke, Mayalen Etcheverry, and Pierre-Yves Oudeyer. "Intrinsically motivated discovery of diverse patterns in self-organizing systems." In: *arXiv preprint arXiv:1908.06663* (2019).

[188] Juliette Renaud. "Évaluer l'utilisabilité, l'utilité et l'acceptabilité d'un outil didactique au cours du processus de conception continuée dans l'usage: Cas d'un outil pour l'enseignement de la lecture de textes documentaires numériques." In: *Éducation et didactique* 2 (2020), pp. 65–84.

[189] Britta Renner. "Curiosity about people: The development of a social curiosity measure in adults." In: *Journal of personality assessment* 87.3 (2006), pp. 305–316.

[190] C.M Roebers, N. von der Linden, and P. Howie. "Favourable and unfavourable conditions for children's confidence judgments." In: *British Journal of Developmental Psychology* 25 (2007), pp. 109–134. DOI: https://doi.org/10.1348/026151006X104392.

[191] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. "Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system." In: *Learning and instruction* 21.2 (2011), pp. 267–280.

[192] Samuel Ronfard, Imac M Zambrana, Tone K Hermansen, and Deborah Kelemen. "Question-asking in childhood: A review of the literature and a framework for understanding its development." In: *Developmental Review* 49 (2018), pp. 101–120.

[193] Leon Rozenblit. "The misunderstood limits of folk science: An illusion of explanatory depth." In: *Cognitive Science* 26 (Oct. 2002), pp. 521–562. DOI: 10.1016/S0364-0213(02)00078-2.

[194] Azzurra Ruggeri and Tania Lombrozo. "Children adapt their questions to achieve efficient search." In: *Cognition* 143 (2015), pp. 203–216.

[195] Azzurra Ruggeri, Caren M Walker, Tania Lombrozo, and Alison Gopnik. "How to help young children ask better questions?" In: *Frontiers in psychology* 11 (2021), p. 586819.

[196] Mark A Runco and Shawn M Okuda. "Problem discovery, divergent thinking, and the creative process." In: *Journal of youth and adolescence* 17.3 (1988), pp. 211–220.

[197] Richard M. Ryan and Edward L. Deci. *Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being*. US, 2000. DOI: 10.1037/0003-066X.55.1.68.

[198] Richard M Ryan and Edward L Deci. "Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions." In: *Contemporary educational psychology* 61 (2020), p. 101860.

[199] Mark A Sabbagh and Dare A Baldwin. "Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development." In: *Child development* 72.4 (2001), pp. 1054–1070.

[200] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. "Statistical learning by 8-month-old infants." In: *Science* 274.5294 (1996), pp. 1926–1928.

[201] Angela K Salmon and Maria Ximena Barrera. "Intentional questioning to promote thinking and learning." In: *Thinking Skills and Creativity* 40 (2021), p. 100822.

[202] Simone Schaeffner, Nicolas Chevalier, Maki Kubota, and Julia Karbach. "Metacognitive training." In: *Cognitive training: An overview of features and applications* (2021), pp. 255–270.

[203] Ulrich Schiefele, Andreas Krapp, and Adolf Winteler. "Interest as a predictor of academic achievement: A meta-analysis of research." In: (1992).

[204] Wolfgang Schneider and Kathrin Lockl. "Procedural metacognition in children: Evidence for developmental trends." In: *Handbook of metamemory and memory* 14 (2008), pp. 391–409.

[205] Laura E Schulz and Elizabeth Baraff Bonawitz. "Serious fun: preschoolers engage in more exploratory play when evidence is confounded." In: *Developmental psychology* 43.4 (2007), p. 1045.

[206] Nicholas Shea, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D Frith. "Supra-personal cognitive control and metacognition." In: *Trends in cognitive sciences* 18.4 (2014), pp. 186–193.

[207] M. D. Shermis and J. Burstein. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge/Taylor& Francis Group, 2013.

[208] A. Silvervarg Flycht-Eriksson and Arne Jönsson. "Towards a conversational pedagogical agent capable of affecting attitudes and self-efficacy." In: *Proceedings of the Second Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity*. 2010.

[209] Paul J Silvia, Beate P Winterstein, John T Willse, Christopher M Barona, Joshua T Cram, Karl I Hess, Jenna L Martinez, and Crystal A Richard. "Assessing creativity with divergent thinking tasks: exploring the reliability and validity of new subjective scoring methods." In: *Psychology of Aesthetics, Creativity, and the Arts* 2.2 (2008), p. 68.

[210] H Simon. "The cat that curiosity couldn't kill." In: *Work-ing Paper, Department of Psychology, Carnegie Mellon University, Pittsburgh* (1992).

[211] Yadurshana Sivashankar, Myra Fernandes, Pierre-Yves Oudeyer, and Hélène Sauzéon. "The beneficial role of curiosity on route memory in children." In: *Frontiers in Cognition* 3 (2024), p. 1346280.

[212] Burrhus F Skinner. "The shame of American education." In: *American Psychologist* 39.9 (1984), p. 947.

[213] Beate Sodian and Uta Frith. "Metacognition, Theory of Mind, and Self-Control: The Relevance of High-Level Cognitive Processes in Development, Neuroscience, and Education." In: *Mind, Brain, and Education* 2.3 (2008), pp. 111–113.

[214] Rayne A Sperling, Bruce C Howard, Lee Ann Miller, and Cheryl Murphy. "Measures of children's knowledge and regulation of cognition." In: *Contemporary educational psychology* 27.1 (2002), pp. 51–79.

[215] Aimee E Stahl and Lisa Feigenson. "Observing the unexpected enhances infants' learning and exploration." In: *Science* 348.6230 (2015), pp. 91–94.

[216] Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. "I Do Not Understand What I Cannot Define: Automatic Question Generation With Pedagogically-Driven Content Selection." In: *ArXiv* abs/2110.04123 (2021).

[217] Md Arafat Sultan, Steven Bethard, and Tamara Sumner. "Towards automatic identification of core concepts in educational resources." In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (Dec. 2014), pp. 379–388. DOI: 10.1109/JCDL.2014.6970194.

[218] Alexandr Ten, Pramod Kaushik, Pierre-Yves Oudeyer, and Jacqueline Gottlieb. "Humans monitor learning progress in curiosity-driven exploration." In: *Nature communications* 12.1 (2021), p. 5972.

[219] Bill Thompson and Thomas L Griffiths. "Human biases limit cumulative innovation." In: *Proceedings of the Royal Society B* 288.1946 (2021), p. 20202752.

[220] Steve Tipps. "Making Better Guesses : A goal in Early childhood Science." In: 82(1) (1982), pp. 29–37. DOI: 10.1111/j.1949-8594.1982.tb17158.x.

[221] Edward C Tolman. "Cognitive maps in rats and men." In: *Psychological review* 55.4 (1948), p. 189.

[222] Jonathan G Tullis and Aaron S Benjamin. "Cueing others' memories." In: *Memory & cognition* 143.4 (2015), pp. 634–646.

[223] R. J. Vallerand, M. R. Blais, N. M. Brière, and L. G. Pelletier. "Construction et validation de l'échelle de motivation en éducation (EME) [Construction and validation of the Motivation toward Education Scale]." In: *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement* 21 (1989), pp. 323–349. URL: https://doi.org/10.1037/h0079855.

[224] Marcel VJ Veenman, Bernadette HAM Van Hout-Wolters, and Peter Afflerbach. "Metacognition and learning: Conceptual and methodological considerations." In: *Metacognition and learning* 1 (2006), pp. 3–14.

[225] Lev S Vygotsky. *Mind in society (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.)* 1978.

[226] Shirlene Wade and Celeste Kidd. "The role of prior knowledge and curiosity in learning." In: *Psychonomic bulletin & review* 26 (2019), pp. 1377–1387.

[227] M Fernanda Wagstaff, Gabriela L Flores, Rawia Ahmed, and Sarah Villanueva. "Measures of curiosity: A literature review." In: *Human Resource Development Quarterly* 32.3 (2021), pp. 363–389.

[228] Jennifer L Weible and Heather Toomey Zimmerman. "Science curiosity in learning environments: developing an attitudinal scale for research in schools, homes, museums, and the community." In: *International Journal of Science Education* 38.8 (2016), pp. 1235–1255.

[229] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding." In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023, pp. 75–78. URL: https://dl.acm.org/doi/abs/10.1145/3581754.3584136.

[230] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding." In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023, pp. 75–78.

[231] Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Hélène Sauzéon, and Pierre-Yves Oudeyer. “Selecting Better Samples from Pre-trained LLMs: A Case Study on Question Generation.” In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 12952–12965. URL: https://aclanthology.org/2023.findings-acl.820.

[232] Philip David Zelazo and Stephanie M Carlson. “Hot and cool executive function in childhood and adolescence: Development and plasticity.” In: *Child development perspectives* 6.4 (2012), pp. 354–360.

[233] Xiaoming Zhai. “ChatGPT user experience: Implications for education.” In: *Available at SSRN 4312418* (2022).

[234] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. “Calibrate before use: Improving few-shot performance of language models.” In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12697–12706.