

Toward More Customized, Differentiated, and Reflective Learning Environments through Human-AI Collaboration: Design and Evaluation of Analytics-Driven Teacher Support Tools

Kexin “Bella” Yang

CMU-HCII-25-104

September 2025

Ph.D. Thesis

Human-Computer Interaction Institute,
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
kexinyang.bella@gmail.com

Thesis Committee:

Dr. Vincent Aleven (Human-Computer Interaction Institute, CMU), Co-chair

Dr. Nikol Rummel (Educational Psychology and Technology, Ruhr University Bochum & Human-Computer Interaction Institute, CMU), Co-chair

Dr. Kenneth Holstein (Human-Computer Interaction Institute, Carnegie Mellon University)

Dr. Xu Wang (Computer Science and Engineering, University of Michigan)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright ©2025 Kexin “Bella” Yang. All rights reserved.

The research reported here was supported, in whole or in part, by the National Science Foundation (NSF) under Grant #1822861 and #2119501 to Carnegie Mellon University. All views expressed in this document are those of the author and do not necessarily reflect the views of the sponsoring agencies.

PREFACE

This dissertation was written in Pittsburgh during 2024–2025, marking the culmination of my Ph.D. journey at Carnegie Mellon University.

My initial interest in AI in education dates back to the end of my undergraduate studies at Beijing Normal University, when I volunteered to teach in rural Guizhou, China. That experience showed me what education is like in underdeveloped areas, where resources are scarce. However, I also saw the immense potential for education technology to be transformative. More importantly, it taught me that teachers must be supported and equipped with the proper scaffolding to benefit from educational technologies fully.

My path toward this field deepened when I completed my master's degree at CMU and first encountered the world of Artificial Intelligence in Education (AIED). The turning point came when I enrolled in the course *Personalized Online Learning*, taught by my advisor, Prof. Vincent Aleven. In that class, I witnessed a demonstration of Lumilo, the research project of Prof. Kenneth Holstein (who later became my mentor and committee member). The project fascinated me, introduced me to Ken, and sparked my enduring interest in AIED.

My first research project (Chapter 3.1) was also my first authentic taste of academic research in AIED and HCI. It was a magical experience—something new and exhilarating. I still remember the moment my first CSCW paper was accepted, after several rounds of rejection and revision. The feeling of leading my own project and having ownership over it was empowering.

During my Ph.D., I was fortunate to work in a dynamic environment where creativity and cutting-edge research were encouraged. I collaborated not only with outstanding researchers, but also with dedicated teachers and students in the greater Pittsburgh area. Through field studies, I learned firsthand how AI might work in K–12 classrooms. These experiences taught me the importance of designing with pragmatic classroom constraints. It was deeply rewarding to see teachers and students using the technologies we built, and to witness the dedication teachers bring to supporting their students' learning.

This dissertation reflects my hope to make learning environments more effective for both students and teachers—by empowering educators to customize technologies and materials, enabling more personalized and dynamically differentiated learning, and supporting teachers in reflecting on and improving their practice.

KEYWORDS.

Artificial Intelligence in Education (AIED), Human-AI Collaboration, Human-AI Co-orchestration, Learning Analytics, Multimodal Learning Analytics (MMLA), Educational Technology, Intelligent Tutoring Systems (ITS), Teacher Support Tools, Customization, Differentiated Learning, Dynamic Differentiation, Teacher Reflection, Teacher Agency, Student Agency, Classroom Orchestration, Computer-Supported Collaborative Learning (CSCL), Individual Learning, Fluid Social Transitions, Teacher Professional Development, Crowdsourcing, Data Simulation, Transaction Log Data, User-Centered Design, Participatory Design, Prototyping Study, Controlled Experiment, Teacher Surveys, Field Studies, K–12 Classrooms, Middle School Mathematics, Teacher Preferences, Student Engagement, Classroom Dynamics, Teacher Location Tracking.

To my parents, mentors and friends, without whom this thesis would not be possible.

ACKNOWLEDGEMENTS

A PhD is a unique journey for everyone who experiences it. I have always been moved and humbled by the help and support that make the completion of a PhD thesis possible. I put my heart into writing this acknowledgement, with a sense of nostalgia as I look back on the past five years. The PhD journey has shaped me profoundly, yet I have never taken it for granted. I thank the PhD admissions committee for granting me the privilege of joining this institute. Throughout all five years, I have spared no effort in striving to conduct good research. “Don’t settle for mediocre—strive for excellence.” If there is one guiding sentence I carry from my PhD, that is it. There are countless moments I remember dearly. For example, I recall when I first started my research project—the long weeks of iteration and refinement of research questions, and the patience of my first mentors, Prof. Kenneth Holstein and Prof. Vincent Alevan, who first introduced me to the world of AIED and HCI research.

First, I would like to thank my PhD advisors and committee co-chairs. Vincent Alevan, who trusted me to lead a two-year research project during the time when I was a master’s student, and dedicated countless hours to brainstorming and discussion. He truly loves his career, the field of AIED, and the pursuit of great research. His passion and pursuit of excellence have influenced me profoundly. His feedback has always been superb, with great attention to detail. Prof. Nikol Rummel has shown me the importance of project management, rigor, organization, and a relentless drive for excellence. Together, my co-chairs have shaped how I think and how I conduct research. They taught me perseverance and grounded me when the ups and downs of a PhD felt overwhelming. Five years have passed as quickly as a single day, marked each week by our one-hour research meetings. Their commitment, dedication, and strong work ethic have been among the most inspiring examples I have ever witnessed.

I am also deeply thankful to my committee members, Xu Wang and Ken Holstein. I especially thank them for their support during my early PhD years—and even before I began—when I knew little about what this journey entailed and they had already experienced it. They always treated me with patience, kindness, and generosity, sharing not only their research knowledge but also their perspectives on life and tips for navigating a PhD, validating me whenever I felt uncertain. I hope to have further opportunities to collaborate with them in the future.

I am grateful for the internship opportunities that allowed me to explore applications of large language models in education and creative workflows—at ACT Inc. (as an AI research intern) and Autodesk Inc. (in the HCI and Visualization group). I owe great thanks to my

mentors during these internships: Justin Matejka, Jo Vermeulen, John Thompson, Sungjin Nam, Scott Wood, and Yuchi Huang.

Over the years, I have witnessed the growth of the CATS Lab, and I am immensely thankful to have worked alongside its current and past members: Naomie Williams, Conrad Borchers, Tomohiro Nagashima, Kenneth Holstein, Jonathan Sewall, Octav Popescu, Qiao Jin, Meng Xia, LuEttaMae Lawrence, Ahmed Ashraf Butt, Shamyia Karumbaiah, Luisa Wellert, Max Benson, Qianru Lyu, Ashish Gurung, Jionghao Lin, Yun Huang, and others.

As a junior researcher, teaching assistant, and collaborator, I have greatly enjoyed working with colleagues around the world, including: Conrad Borchers, Ann-Christin Falhs, Vanessa Echeverria, Shamyia Karumbaiah, Xu Wang, Meredith Thompson, Tomohiro Nagashima, Joseph Jay Williams, Carolyn Rosé, Jionghao Lin, Kenneth R. Koedinger, LuEttaMae Lawrence, Zijing Lu, Hongyu Mao, Boyuan Guo, Xuejian Wang, Junhui Yao, Jessica Ceha, Xiaofei Zhou, David Lindlbauer, Raelin Musuraca, Motahhare Eslami, Shiyan Jiang, Bruce McLaren, Ashish Gurung, and Cindy Liu. I am also grateful to the many researchers I met and exchanged feedback with at various HCI and AIED conferences.

I would also like to thank the research assistants and interns I had the pleasure of working with: Mandy Pan, Manqing Yu, Maria Ma, Yikuan Fang, Ashley Xu, Hongyu Mao, Zijing Lu, Jinjin Du, Jori Blankestijn, Megan Lin, Yvonne Yi Zhao, Menghan Liu, Maggie Hung, Yu Fang, Sijia Fan, Sofia Wheeler, Sahityasree Subramanian, Subhashini Ramkumar, Alexander Liu, Xun Qin, and Donovan So.

Before my PhD, I was fortunate to receive mentorship and guidance from Prof. Edith Law during my year at the University of Waterloo in the HCI Lab, and from Prof. Carolyn Rosé and Prof. Shiyan Jiang on automating writing feedback with NLP at the Language Technologies Institute at CMU. I also thank all METALS faculty, including Sharon Carver, Lauren Herckis, Bruce M. McLaren, Ken Koedinger, Marti Louw and many others, for equipping me with the knowledge that paved the way for my PhD. I fondly remember hugging the core faculty members with deep gratitude at Prof. Ken Koedinger's home during graduation. I also thank the HCII master's and PhD administrative staff—Becky Wang, Bill Carney, Eric Davidson, Queenie Kravitz, Geoff Kaufman, Jo Bodnar, and Michael Bett—for their invaluable support.

On the personal side, I thank my parents, Suling and Yibin, for their unwavering support. I also thank the many friends and colleagues who have given me encouragement, advice, company, inspiration, and care, including (in no particular order): Yurun Tian, Xinyue Chen, Yixin Chen, Hongyu Mao, Yunzhi Li, Tong Wang, Franklin Mingzhe Li, Yiwen Song, Harley Chang, Canhai Chen, Xiaofei Zhou, Yihong Shi, Yuqing Xie, Wei Yang, Sherry Tongshuang

Wu, Xieyang Michael Liu, Meng Xia, Xinyu (Scinko) Yang, Hyunsun Cho, Jiayi Zhou, Hao Li, Wenjie Li, Zhixuan (Beth) Bai, Claire Yue Wang, Liwei Ariel Liu, Shaojie Bai, Yewen Fan, Xueyan (Victoria) Zou, Peiling Jiang, Yankun (Alex) Zhao, Huy Nguyen, Naiyu Fang, Shuailin (Glen) Pan, Yihang Shen, Lantian Xu, Yuze Liao, Xiangyu Zhang, Jingyu Tang, Toby Li, Yihhsing Liu, Lei Wang, Xin Hong, Jing Lin, and Xiyao Li.

Finally, I would like to acknowledge myself—for staying resilient through the ups and downs, and for holding on to this journey with perseverance. The PhD has taken me to places I never imagined, given me experiences I never expected, and allowed me to meet people I am deeply grateful to have known. To everyone not explicitly listed here, thank you for your support before, during, and hopefully after my PhD. As I embark on a new journey, I know that completing this doctoral thesis is only the beginning of an endless exploration of life and the world.

ABSTRACT

Classroom teaching with AI-based educational software involves complex tasks spanning across three phases: before class, during class, and after class. 1) **Before class**, teachers need to do adequate lesson planning and preparation. It is important to ensure that the instructional materials and technology to leverage are *customized* to the classroom context and aligned with the curriculum. 2) **During class**, teachers frequently improvise and constantly differentiate their actions to serve the needs of their different students. 3) **After class ends**, post-class *reflection* can help teachers to think critically about what happened, to flexibly adapt their instructions. It is important to provide teacher support in all phases. The advance of educational technology and multi-modal learning analytics (MMLA) (Cukurova et al., 2020; Escueta et al., 2017; Holstein et al., 2019b; Holstein & Aleven, 2022) holds promise to support teachers and students to achieve more effective teaching and learning, by empowering teachers to focus on instructional tasks that they do the best and that they enjoy doing. My PhD research aims to create synergy between humans and AI as we design, develop, and evaluate educational technology that fosters more **customized, differentiated, reflective learning environments**.

Customization of Instructional Materials and Technology. While there are many ways to customize learning environments, my work focuses on instructional materials and technology. Prior literature has noted the need and importance of aligning *instructional materials and technology* with teachers' practice and preferences to achieve sustained use and adoption.

Customization of instructional materials. For instructional materials, I studied how to achieve *customized* instructional materials (mathematical hints) while saving time for domain experts through a crowdsourcing pipeline. An important question left open by past work is how to support crowds with minimal expert guidance in revising and customizing instructional materials. This work found a clear need for customizing hints to individual teachers' preferences and highlights the need for more elaborate scaffolds so the crowd has specific knowledge of the teachers' requirements for hints.

Customization of technology. For technology, I studied how to ensure feasibility (in the sense of student coverage) for dynamic pairing policies (which support pairing up students in a personalized way), in a teacher orchestration tool. Prior work has provided little insight into the feasibility of pairing policies, especially in the ITS context. I studied how to leverage data simulation of historical transaction log data (SimPairing) to assess the feasibility of dynamic learning group formation policies in different classroom contexts. I found that dynamically pairing students based on their in-the-moment wheel-spinning status yields good feasibility for struggling students, even under moderate restrictions on allowed pairings. Importantly,

any fixed instantiation of a policy does not fully account for class variability and extreme cases—policies perform differently across classrooms and sessions, suggesting that policy feasibility may be optimized by gradually loosening parameters or through classroom-level customization. This study contributes to how different dynamic pairing policies have different feasibility results in different classrooms and bridges the literature gap in its investigation of the feasibility of user-centered dynamic pairing policies.

Dynamic Differentiation between Individual and Collaborative Learning. There are many ways to make learning environments support differentiated learning, and my work focuses on dynamically combining different learning modes for students. Prior research found that combining individual and collaborative learning can be more beneficial than either way alone. However, prior work has not created or studied a technology ecosystem to achieve dynamically differentiated learning between individual and collaborative learning and systematically evaluate the value of the dynamic combination of individual and collaborative learning. Grounded in the context of real-time, analytic-based intelligent tutoring systems, my colleagues and I created a technology ecosystem to achieve the dynamic combination of individual and collaborative learning. I then conducted a series of studies to understand teachers' preferences and boundaries in co-orchestrating with AI systems in such dynamic pairing, evaluated in an authentic classroom prototyping study, the feasibility of such dynamic combinations in classrooms, and conducted a controlled study to evaluate how dynamic transitions are compared with traditional transitions.

I discovered teachers' preferences about the respective roles of teachers, students, and AI in a co-orchestration system of dynamic pairing. I found that dynamic transitions between different activities have pedagogical value in the eyes of teachers and, to a lesser degree, students. Supporting dynamic transitions in actual classrooms is worthwhile from a pedagogical perspective, yet not without challenges, such as finding the optimal timing, content, and partner during such transitions. I found that teachers and students demonstrated a subjective preference for the dynamic condition, which stemmed from its flexibility and individualization. However, for some of the test items, the standard condition had more learning gains than the dynamic condition.

Multi-modal Based Teacher Reflection Tool. My work focuses on supporting teacher reflection through multimodal learning analytics (MMLA) in classrooms where students use ITS. Teacher reflection can be an important component in advancing the quality of instruction. Teacher reflection holds potential to scaffold teachers' critical thinking (Korthagen, 2004), support knowledge construction in teaching (Conway, 2001), and promote self-regulation in teachers (Boud & Falchikov, 2007; Singh, 2008). However, more work is needed to understand how a reflection tool can be designed to fit into teachers' routines and keep in mind their preferences for data sharing and privacy considerations. With the help of

colleagues, I studied teachers' preferences about multi-modal data analytics collection and analytics use in reflection tools. I found teachers showed the strongest interest in data about motivating students and interactions with students (teacher analytics) and students' learning and progress (student analytics). I also learned about specific teachers' data collection and sharing preferences in MMLA.

In my final study related to Reflecto, a tool that leveraged multi-modal data analytics (including location sensing and transaction log data) to support teacher reflection, I started with refining and pilot testing this teacher reflection tool, and a small-scale prototyping activity. Then I led an exploratory classroom study with actual teachers and students, which addressed the following research questions: 1) RQ1: How do teachers use the MMLA-based teacher reflection tool during individual and collaborative reflection sessions? 2) RQ2: How does Reflecto influence teachers' understanding of their teaching practices and future intentions for interacting with students? 3) RQ3: How does the use of an MMLA-based reflection tool impact teachers' actual teaching practices? 4) RQ4: What are teachers' perceptions of the MMLA-based reflection tool, and what design feedback do they offer?

My PhD research leverages various methodologies, including pipeline and tool design, data simulation and mining, user-centered research, as well as both prototyping study and controlled classroom experiments. My research focuses on aiding teachers in *customizing* instructional materials and technology (i.e., pairing algorithms) and facilitating *dynamically-differentiated* classrooms, as well as reflective instructional practices using Multimodal Learning Analytics (MMLA). This dissertation advances knowledge on **customization of instructional materials, orchestration technology, and teacher reflection** in AI-based learning environments. I demonstrate the promise and value of leveraging minimal expert guidance and a crowdworker pipeline to save teachers' time in instructional materials customization. My work also highlights the pedagogical value of dynamically combining individual and collaborative learning through a technology ecosystem, as well as the nuances and design insights related to human-AI shared control in this context. In addition, I show that an MMLA-based reflection tool (leveraging teacher location and student log data from ITSS) can help teachers validate existing understandings of classroom practice, challenge assumptions about their students and classrooms, and generate actionable insights.

Table of contents

Preface.....	1
Acknowledgements.....	3
KeyWords.....	7
Abstract.....	8
Chapter 1 Introduction.....	13
1.1 Customization.....	15
1.2 Dynamic Differentiation.....	17
1.3 Reflection.....	18
Chapter 2 Theoretical Background.....	24
2.1 Customization in the Education Field.....	25
2.2. Human-AI Co-orchestration in the Classroom.....	27
2.3 Dynamically Combining Individual and Collaborative Learning.....	28
2.4 MMLA used to support reflective teaching environments.....	30
I. Customization.....	32
Chapter 3 Customized Instructional Materials and Technology.....	33
3.1 Pipeline Design for Teacher Customization of Materials.....	33
3.2 Customizing Dynamic Group Formation Policy for Classroom Context.....	44
II. Differentiation.....	65
Chapter 4 User Research on Teacher Preference and Boundary in Human-AI control (in Context of Dynamically Differentiated Learning between Individual and Collaborative Learning).....	66
4.1 Study Summary.....	67
4.2 Background and Motivation.....	68
4.3 Methods and Study Design.....	69
4.4 Results.....	72
4.5 Discussion.....	75
4.6 Design Challenges and Directions.....	76
4.7 Study Contribution and How it Connects to the Thesis Theme.....	77
Chapter 5 Pair-Up: Supporting Dynamic Combination of Individual and Collaborative Learning.....	78
5.1 Study Motivation and Context.....	80
5.2 Research Questions.....	82
5.3 Participants and Procedure of Classroom Prototyping.....	83
5.4 Data Collection and Analysis.....	84
5.5 Results: Students' Learning and Teachers' Orchestration Process During Dynamic Transitions (RQ1 and RQ2).....	87
5.6 Results: Teachers' and Students' Perceived Value and Design Opportunities (RQ3 and RQ4)...	93

5.7 Learning and Orchestration Process of Dynamic Transitions, and New Classroom Dynamics..	97
5.8 Experiences with and Perceived Value of Dynamic Transitions.....	98
5.9 Design Implications and Challenges.....	99
5.10 Ethical Implications.....	101
5.11 Conclusion and Role in Thesis.....	101
Chapter 6. A Controlled Experiment Comparing Dynamic and Standard Transitions	102
6.1 Research Questions.....	104
6.2 High-level Results.....	104
III. Reflection.....	110
Chapter 7 Teacher Preference in Building an MMLA Reflection Tool.....	111
7.1 Background and Motivation.....	113
7.2 Methods.....	114
7.3 Results.....	116
7.4 Discussion and Study Takeaway.....	121
Chapter 8 Reflecto - Leveraging Multimodal Learning Analytics in Teacher Reflection Tool.....	124
8.1 Reflecto - a teacher-facing reflection tool based on multi-modal learning analytics.....	125
8.2 User Testing and Prototyping of the Reflection Tool: Reflecto.....	129
Chapter 9 Reflecto Classroom Exploratory Study (Motivation and Methodology).....	135
9.1 Goal of Evaluation Study.....	135
9.2 Research Questions.....	135
9.3 Study Procedure.....	136
9.4 Iteration to Reflection Activity Protocol and Interview Protocols.....	142
9.5 Data Analysis.....	144
Chapter 10 Reflecto Classroom Exploratory Study (Results and Discussion).....	148
10.1 Results for Research Question 1.....	148
10.2 Results for Research Question 2.....	154
10.3 Results for Research Question 3.....	167
10.4 Study Results for Research Question 4.....	170
10.5 Study Results Summary and Discussion.....	179
10.6 Study Conclusion.....	187
Chapter 11 Summary and Thesis Contribution.....	187
11.1 Customization of Instructional Materials and Technology.....	187
11.2 Dynamic Differentiation between Individual and Collaborative Learning.....	189
11.3 MMLA-based Teacher Reflection Tool.....	190
11.4 Conclusion.....	191
Reference.....	192

CHAPTER 1 INTRODUCTION

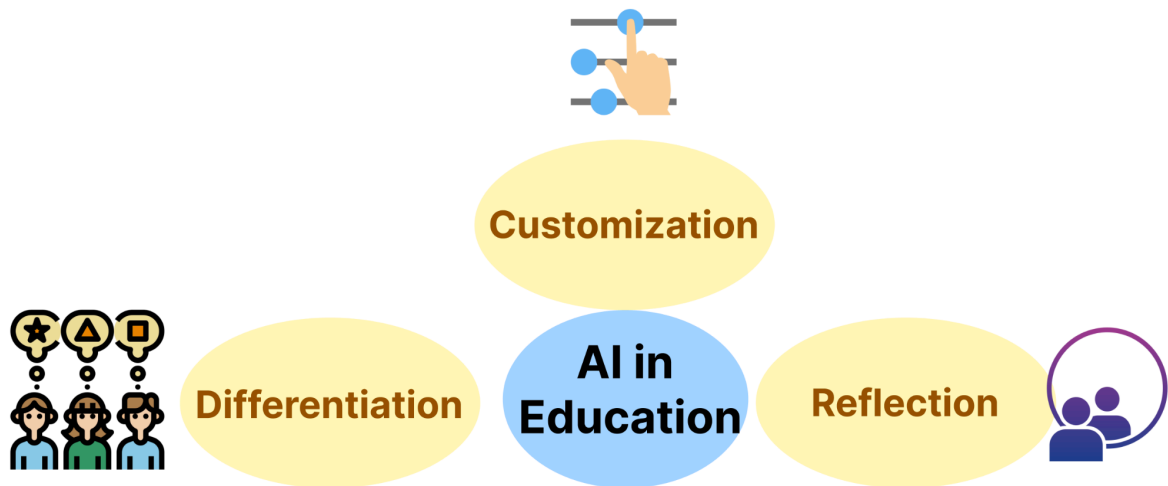


Fig. 1.1. Main Thesis Research Theme: How Technologies and AI-Based Tools and Analytics Can Support Teachers and Create Customized, Differentiated, and Reflective Educational Environments

Classroom teaching is a complex job, as teachers often need to balance instruction, helping students individually, and managing classroom activities (Corno, 2008). AI holds the potential to assist and augment teachers' abilities (Cukurova et al., 2020; Escueta et al., 2017; Holstein et al., 2019b; Holstein & Aleven, 2022). Much work on using AI in education has focused on supporting interactions between technology and students that promote learning (e.g., intelligent tutoring systems) (du Boulay, 2019). It is fair to say that far less work has been done to apply AI to support teachers (du Boulay, 2019).

As du Boulay argued, to maximize the effectiveness of the scarce resource of the teacher's time and attention, training and support are needed to help the teacher decide when to intervene and mediate the interaction between learners and the intelligent tutoring systems (du Boulay, 2019). Through the thoughtful integration of technology, educators can more effectively tailor educational experiences, manage classroom dynamics, and reflect on their teaching methods to improve student outcomes. A unifying question that the thesis aims to address is **how technologies and AI-based tools and analytics can be used to create customized, differentiated educational environments that cater to the dynamic needs of teachers and students and support the teachers to become reflective practitioners** (Schwartz & Schon, 1987) (Fig 1.1).

As claimed in the literature (Kumar, 2015), there are three phases of classroom teaching: 1) Pre-Active Phase, which is the planning phase. 2) Interactive Phase, which refers to the execution of the plan and is also known as the actual classroom teaching. 3) The post-active phase, which is the evaluator phase of teaching, includes tests or quizzes after class ends, observing students' reactions to questions, instructional situations, and comments (Kumar, 2015). **My thesis research aims to make learning environments more customized, dynamically differentiated, and reflective for teachers (Fig. 1.2).** Specifically, this thesis targets topics about 1) how we can achieve **classroom customization** for instructional materials and technology (during the pre-active phase), 2) how we can achieve **dynamically differentiated** collaborative learning (during the interactive phase), and 3) how we can co-design a tool to support teachers' **reflection** (during the post-active phase) (Fig. 1.3).

It is important to note that this coupling of learning environment characterization with the three phases is incidental to my PhD research thesis. These three characteristics are not the only ones that are desirable in each phase. For example, it may also be important for learning activities to be "customized" during class (in the Interactive Phase), but that goes beyond the scope of this thesis. Additionally, some of the characteristics discussed in my thesis may have a common end goal or overlap. For example, customization and differentiation both serve the common purpose of achieving education that can be maximally beneficial for an individual's learning. The main difference is that customization emphasizes stakeholders' agency in making choices or decisions (e.g., teachers requesting which materials or algorithms to use). While differentiation does not have this connotation, as in the case of this thesis, students can engage in differentiated learning without actively being able to choose how the experiences unfold (i.e., the specific choices of how students engage in differentiated learning experiences decided by their teachers).

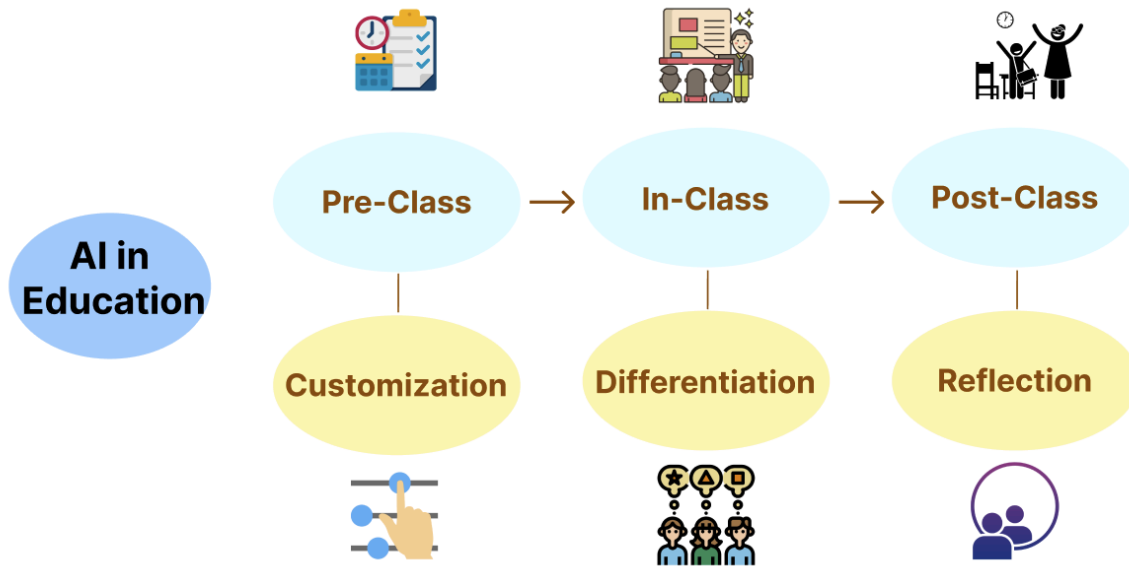


Fig. 1.2. The Three Research Goals for Learning Environments Examined in This Thesis, Aligned With the Three Instructional Phases

1.1 Customization

Definition. In general, customization means that before a product is made, it is “made to fit” the needs of an assigned customer (Schuwer & Kusters, 2014). In my thesis work, I define “customization of AI-based learning technologies” as a process of making adaptations, alterations, or personalizations of the materials, tool features, or functions of AI-based learning technologies. Much prior work has described or studied the landscape of customized learning and education (e.g., (Sharma et al., 2017; Sokolov, 2001). My thesis focuses on two specific forms of customization, namely, customization of instructional materials (mathematical hints in ITS) and customization of the technology (pairing algorithms in this case).

Prior literature has found the need for teachers to create customized instructional materials (Blumenfeld et al., 2000; Matuk et al., 2016; Means et al., 2001). Additionally, prior work found that allowing teachers to customize and revise aspects of AI-based learning software may be essential for wider adoption and for sustaining the relevance of materials in the long term (Blumenfeld et al., 2000; Matuk et al., 2016). However, despite the need, prior work has not found a scalable, time-efficient way to enable stakeholders (teachers and students) to customize instructional materials in AI-based software (Holstein et al., 2017, 2019b; Nye, 2014). To solve this research gap, I leveraged crowdsourcing, asking: **Can we minimize experts’ (teachers’) time and leverage the crowdsourcing pipeline to achieve quality instructional materials customization for teachers?** (Study 1 in Table 1) Prior studies

showed that crowdworkers (MTurkers) can generate instructional materials of reasonable quality for K-12 math when the goal is typically to create instructional resources where none existed previously (e.g., (Aleahmad et al., 2009; Y. Chen et al., 2016)). Less is known about whether crowd workers could succeed in content revision, where they must improve upon existing material. On the one hand, content revision tasks may be easier for crowd workers than generating content from scratch since workers have access to the original materials as a strong starting point for their task, plus any guidance on how to revise the given materials. On the other hand, revision tasks could be harder for crowd workers than content generation tasks, as producing consistent, high-quality revisions requires crowd workers to fully comprehend the original materials and contextualize expert guidance (e.g., brief specs for the revisions), which may not be easy for a non-expert. Revision tasks could be especially hard if the original materials are already of high quality, as in this case. Thus, I conducted a crowdsourcing pipeline and experiment to validate whether the crowd workers were up to the task and contributed results and design insights (Chapter 3.1).

With a similar goal of making the technology (specifically, pairing algorithms) more customized to each classroom context, I simulate the feasibility of different dynamic pairing policies in different classrooms (Study 2 in Table 1). Constructing effective, well-balanced learning groups is an important task in computer-supported collaborative learning (CSCL) (Amara et al., 2016; P. Dillenbourg, 2002; X. Wang et al., 2021). The importance of learning group formation (LGF) has been validated empirically (Huang et al., 2014; Webb et al., 1998). However, **evaluating the feasibility of LGF policies offline prior to implementing them in real classrooms** remains a challenging task due to a lack of readily accessible approaches. To address this problem, I adopt a process I call “*SimPairing*” to simulate pairing policies on authentic data and evaluate their feasibility (Chapter 3.2). In this process, I used transaction data from several classes of students using an ITS collected from classroom studies conducted in U.S. middle schools. I computed and analyzed how the feasibility of several LGF policies changed as each class progressed and how the feasibility varied across different classes. This work reports how different policies may have different feasibility results in different class sessions and considerations about customizing the pairing policies (e.g., parameters) to specific classroom contexts. One novelty of this work is that it studies dynamic pairing policies, considering students’ in-the-moment wheel-spinning status. Also, instead of prescribing *a specific grouping criterion*, *this work envisions that instructors will customize pairing policies and parameters (e.g., the knowledge heterogeneity and similarity) to their classroom contexts, which prior work argued to be especially helpful in the LGF process* (Amara et al., 2016; Echeverria et al., 2020; Maqtary et al., 2019; Yang, Nagashima, et al., 2021). I found that, on average, it is generally feasible to team up struggling students with non-struggling ones, meaning that the groups are sufficiently balanced that pairs can be formed. I found that the policy feasibility in different sessions can be differently influenced

by the required knowledge distance (KD) threshold or ceiling, depending on how closely clustered together students' initial knowledge (IK) is.

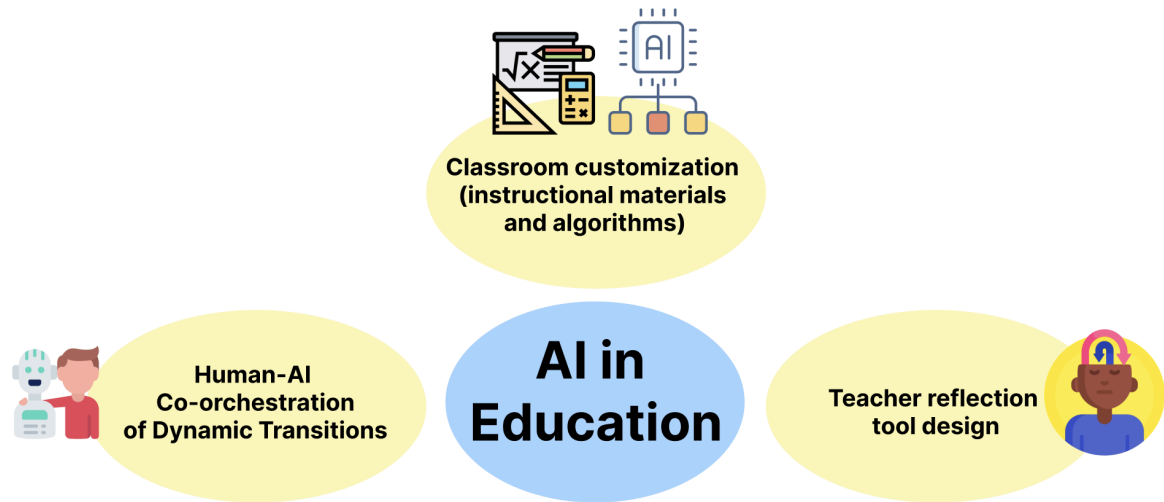


Fig 1.3 Thesis research topics and specific projects and studies relevant to the AI in Education broad themes

1.2 Dynamic Differentiation

Definition: Differentiated instruction is a pedagogical approach that acknowledges and adapts to students' varying talents (Morgan, 2014). This form of teaching is designed to meet the needs of diverse learners in order to alleviate and eliminate disengagement (Morgan, 2014) by emphasizing student responsibility, peer tutoring, flexible grouping, and student choice (Grimes & Stevens, 2009). Differentiation is a complex concept that is difficult to translate from policy into classroom practice, and it requires support for teachers to implement effectively (Mills et al., 2014). My thesis research focuses on one particular type of differentiated learning: transitioning between individual and collaborative learning in a dynamic manner.

Prior research found that combining individual and collaborative learning can be more beneficial than either way alone (J. K. Olsen et al., 2019, 2020). Prior classroom observations found that when multiple students need help, teachers may team students up opportunistically so that students can get help from their peers (Holstein et al., 2019a, 2019b). I refer to such events as *dynamic combinations* of individual and collaborative learning. Dynamic transitions go beyond the pre-planned team formation strategies that are common in classrooms, where most or all students start or stop collaboration at the same time. In dynamic transitions, by contrast, teachers have students transition between learning modes (individual or collaborative) *when the need arises* to be maximally responsive to the fact that students learn at their own pace. *Dynamically combining individual and collaborative learning* has

pedagogical potential but, prior to my study, had yet to be investigated in authentic classrooms. Despite some emerging research and prototypes on transitioning between individual and collaborative learning modes (Echeverria et al., 2020; J. K. Olsen et al., 2020), there was a lack of adequate empirical work that prototyped and evaluated such technologies in classrooms. My research bridged this gap through the participatory design of such an orchestration tool and conducted a classroom study to evaluate the idea of dynamically combining individual and collaborative learning through authentic classroom studies (detailed in Chapters 4, 5, 6). The main questions I tackled are:

- 1) How can we build a technology ecosystem that leverages AI and teachers' strengths, respects teachers' preferences, and supports a dynamic combination of individual and collaborative learning? (Chapter 4)
- 2) Is it feasible for teachers and students to achieve these combinations with technology support? (Chapter 5)
- 3) How do dynamic combinations of individual and collaborative learning compare to traditional pairing in classrooms? (Chapter 6)

Through this thread of research work, I discovered teachers' preferences about the respective roles of teachers, students, and AI in a co-orchestration system of dynamic pairing (e.g., teachers prefer system suggestions and have final control over pairing decisions. I found that dynamic combinations of individual and collaborative learning have pedagogical value in the eyes of teachers and, to a lesser degree, of students. Teachers view dynamic transitions as a way to achieve more differentiated learning that can be better customized to each student's individual pace and preferred learning mode. Supporting dynamic transitions in actual classrooms is worthwhile from a pedagogical perspective, yet not without challenges, such as finding the optimal timing, content, and partner during such transitions.

1.3 Reflection

Prior research found that teacher reflection scaffolds critical thinking, provides a source of knowledge construction in teaching, and promotes self-regulation in teachers (Marcos et al., 2011). Reflection for classroom practice is important to teachers' professional development and ensures flexibility in teaching. Reflection may facilitate teachers' professional development by enhancing teachers' professional learning and knowledge, developing inquiry skills, increasing a greater self-awareness, and leading to change in practice, developing informed action as opposed to following routine teaching action, and developing collaboration among teachers (Çimer et al., 2013). Empirically, some studies have shown teacher reflection to have a positive impact on teachers' practice (Çimer et al., 2013; Cornford, 2002; Gibbons & Farley, 2021). Multi-modal Learning Analytics (MMLA) holds promise to help with teacher reflection, by analyzing data from teachers and students across

digital and physical learning environments (Cukurova et al., 2020). MMLA is defined as “a set of techniques that can be used to collect multiple sources of data in high frequency (video, logs, audio, gestures, biosensors), synchronize and code the data, and examine learning in realistic, ecologically valid, social, mixed-media learning environments” (Giannakos et al., 2022). Given the nascent stage of multimodal data and AI in classrooms, more work is needed to understand how a reflection tool can be designed to fit into teachers’ routines with teachers’ practical needs in mind (Prieto et al., 2020). In my research, I study the following broad questions:

- 1) What analytics on teaching and learning do teachers want to use in reflection practices, and why? (Chapter 7)
- 2) How would teachers like an MMLA-based teacher tool to be designed to support their reflection practices? (Chapter 8)
- 3) How could such an MMLA-based reflection tool impact teachers’ understanding and practices of their classroom through post-class reflection? (Chapter 9 - 10)

This part of the research makes contributions to advance the understanding at the intersection of teacher-reflection and Multimodal Learning analytics (MMLA). I studied how MMLA-based analytics (specifically, teacher location and students’ ITS logs) can be leveraged to build a teacher-reflection tool that is aligned with users’ needs and practices. I also studied how such a tool may support teachers’ reflection, as well as their understanding and actions of their classroom practice.

Summary. In my PhD research, I have conducted studies described in Table 1.1. I delved into various methodologies and thematic areas concerning technology-enhanced learning and teacher support tools. Methodologically, I employed pipeline and tool design, data simulation and mining, and user-centered research. These methods supported the creation of prototypes and conducting both prototyping studies and a controlled experiment. A recurring finding throughout my work is the significance of teacher agency, which can sometimes conflict with the level of agency students prefer. Additionally, teacher preferences often vary based on contextual and nuanced factors, such as student demographics, learning traits, and special educational needs, highlighting the value of user-centered design in educational tools.

Thematic focus was placed on aiding teachers in customizing instructional materials and technology (*customized*) and facilitating dynamic transitions in the classroom (*differentiated*), and effectively reflecting on their teaching (*reflective*).

High-level Organizer: This document is arranged in the following manner: Chapter 1 introduces the thesis as a whole. Chapter 2 introduces the theoretical background, respectively, in customization (2.1), human-AI co-orchestration (2.2), dynamically combining

individual and collaborative learning (2.3), and MMLA use to support reflective learning (2.4). The detailed content in Chapters 3 - 10 can be found in Table 1.1. Chapter 11 describes the high-level contribution the thesis makes.

Table 1.1 Completed Research Studies in the Thesis

Study #	Research Study	
Theme: CUSTOMIZATION		
1	Time	Spring 2019 - Spring 2021
	Study & Methods	Using crowdsourcing to help teacher customization and personalization of instructional materials Methods: Pipeline Design and Experimental Evaluation.
	Chapter	Chapter 3.1
	Publication	Kexin Bella Yang , Tomohiro Nagashima, Junhui Yao, Joseph Jay Williams, Kenneth Holstein, Vincent Aleven. Can crowds customize instructional materials with minimal expert guidance? Exploring teacher-guided crowdsourcing for improving hints in an AI-based tutor. In <i>Proceedings of the ACM on Human-Computer Interaction (CSCW’ 21, <u>PDF</u>)</i>
2	Time	Fall 2020 - Spring 2021
	Study & Methods	SimPairing: Studying feasibility for dynamic pairing policies through log data simulation Methods: Data Mining and Simulation.
	Chapter	Chapter 3.2
	Publication	Kexin Bella Yang , Vanessa Echeverria, Xuejian Wang, LuEttaMae Lawrence, Kenneth Holstein, Nikol Rummel, Vincent Aleven. (2021) Exploring Policies for Dynamically Teaming Up Students through Log Data Simulation. In <i>Proceedings of the 14th International Conference on Educational Data Mining. (EDM’ 21, <u>PDF</u>)</i>
Theme: DIFFERENTIATION		
3	Time	Fall 2020 - Spring 2021
	Study & Methods	I designed storyboards of possible scenarios of human-AI co-orchestration of dynamic pairing, and surveyed 54 teachers on their preferences. Methods: Human-centered design-based research.
	Chapter	Chapter 4
	Publication	Kexin Bella Yang , LuEttaMae Lawrence, Vanessa Echeverria, Boyuan Guo, Nikol Rummel, Vincent Aleven. “Surveying Teachers’ Preferences and Boundaries

		regarding Human-AI Control in Dynamic Pairing of Students for Collaborative Learning”. (<i>EC-TEL’ 21</i> , PDF)
4	<i>Time</i>	Fall 2020 - Spring 2022
	<i>Study & Methods</i>	Co-design and Participatory design for the Pair-Up Tool and its Technology Ecosystem Methods: Human-centered design-based research.
	<i>Chapter</i>	Chapter 5: The first two paragraphs before 5.1 describe the Pair-Up tool and the technology ecosystem.
	<i>Publication</i>	• Kexin Bella Yang, Zijing Lu, Vanessa Echeverria, Jonathan Sewall, Luettamae Lawrence, Nikol Rummel, Vincent Aleven. “Technology Ecosystem for Orchestrating Dynamic Transitions between Individual and Collaborative AI-tutored Problem Solving”. (<i>AIED’ 22</i> short paper, PDF) • LuEttaMae Lawrence, Boyuan Guo, Kexin Bella Yang, Vanessa Echeverria, Jimmy Kang, Vikrant Bathala, Christina Li, William Huang, Vincent Aleven, Nikol Rummel “Co-Designing AI-Based Orchestration Tools to Support Dynamic Transitions: Design Narratives Through Conjecture Mapping” (ICLS’ 22, PDF., best design paper) • LuEttaMae Lawrence, Vanessa Echeverria, Kexin Bella Yang, Vincent Aleven, Nikol Rummel. How teachers conceptualise shared control with an AI co-orchestration tool: A multiyear teacher-centred design process. (August 2023 , British Journal of Educational Technology) https://doi.org/10.1111/bjet.13372
5	<i>Time</i>	Fall 2021 - Spring 2023
	<i>Study & Methods</i>	Prototyping study for feasibility of dynamic transitions in authentic K-12 classrooms Methods: Prototyping classroom field study
	<i>Chapter</i>	Chapter 5
	<i>Publication</i>	Kexin Bella Yang , Vanessa Echeverria, Zijing Lu, Hongyu Mao, Kenneth Holstein, Nikol Rummel, Vincent Aleven. “Pair-Up: Prototyping Human-AI Co-orchestration of Dynamic Transitions between Individual and Collaborative Learning in the Classroom”. Proceedings of the 2023 CHI conference on human factors in computing systems (<i>CHI’ 23</i> , PDF)
6	<i>Time</i>	Fall 2022 - Fall 2023 (Study time) 2024-2025 for Paper Revision and Resubmission
	<i>Study & Methods</i>	Comparing Dynamic Transitions in Authentic K-12 Classrooms Methods: Controlled Experimental Classroom Study.
	<i>Chapter</i>	Chapter 6
	<i>Publication</i>	Comparing Dynamic versus Pre-planned Combinations of Individual and Collaborative Learning in a Classroom Experiment. Yang et al, in revision.

Theme: REFLECTION		
7	<i>Time</i>	Fall 2023 - Fall 2024
	<i>Study & Methods</i>	Surveying Teachers' Preference in Building MMLA Reflection Tool Methods: Human-Centered Design Methods.
	<i>Chapter</i>	Chapter 7
	<i>Publication</i>	Kexin Bella Yang*, Conrad Borchers*, Ann-Christin Falhs, Vanessa Echeverria, Shamyia Karumbaiah, Nikol Rummel, Vincent Aleven; Leveraging Multimodal Classroom Data for Teacher Reflection: Teachers' Preferences, Practices, and Privacy Considerations. (EC-TEL' 24, Full paper)
8	<i>Time</i>	Fall 2024 - Fall 2025
	<i>Study & Methods</i>	MMLA-based Teacher Reflection Tool and its Evaluation in Classroom Methods: Human-Centered Design, Classroom Field Study.
	<i>Chapter</i>	Chapter 8 describes the tool Reflecto, and its testing and prototyping process. Chapter 9 describes the methods and procedure for the exploratory study of Reflecto. Chapter 10 describes the results for the exploratory study of Reflecto.
	<i>Publication</i>	Kexin Bella Yang, Yi Zhao, Conrad Borchers, Yu Fang, Jasmin Kim, Ahmed Butt, Octav Popescu, Sahil Gupte, Ariel Chiu, Peggy Shen, Nikol Rummel, Vincent Aleven. Reflecto: A Teacher Reflection Tool Leveraging Multimodal Learning and Teaching Analytics (AIED'25, Interactive Event)

CHAPTER 2 THEORETICAL BACKGROUND

In this chapter, I discuss related literature, respectively, about customization in the education field (2.1), human-AI co-orchestration in the classroom (2.2), dynamically combining individual and collaborative learning (2.3), and MMLA use to support reflection (2.4). The reason for this division is that these are major themes relevant to my thesis. The tool design in my thesis often needs to consider how customization plays a role in sustained technology adoption (2.1), which is particularly relevant to my previous work on customization of instructional materials and technology (Chapter 3). Human-AI perspective and human preferences in such co-orchestration (2.2) and dynamically combining individual and collaborative learning (2.3) are relevant to the “differentiation” theme in my thesis work. We combined these two strands into a technology ecosystem that can allow for dynamically differentiated learning (in the form of combining individual and collaborative learning in a self-paced manner). This technology ecosystem is featured by a core, new component, Pair-Up, a co-orchestration teacher tool. Thus, I discussed prior work on the co-orchestration (2.2) as well as the potential benefits of combining individual and collaborative learning (2.3). Lastly, connecting to the “reflective” theme, I discussed the prior literature on teacher reflection, and the potential to leverage MMLA to education (especially, reflection) in 2.4.

2.1 Customization in the Education Field

Two strands of my work relate to customization for AI-based educational software: 1) leveraging crowdsourcing for instructional materials customization (Chapter 3.1); 2) dynamic learning group formation policies to evaluate feasibility in the classroom (Chapter 3.2). I discuss relevant past research and the interesting questions past work has left open that are addressed in the dissertation.

Providing customized educational experiences (Sharma et al., 2017; Sokolov, 2001) and having customized instructional materials for teachers is important for sustaining teachers’ use (Blumenfeld et al., 2000; Matuk et al., 2016; Means et al., 2001). Many researchers aim to support the customization of educational tools. For example, to support software engineering education, researchers found that children (grades 3-4) prefer a game with a cosmetically customizable environment, where the *students* can make choices about a game’s visual or audio attributes without directly affecting the content of the gameplay over its non-customizable counterpart. Such preferences only happen when they can contrast these two versions (not when they cannot contrast them), although they found children enjoy learning and learn equally from customizable and non-customizable games (Javora et al., 2021). Some research has looked into using AI methods (e.g., ensemble learning algorithms) to provide customized learning paths for students (Smirani et al., 2022).

Customization of AI-based Educational Software. AI-based educational software is increasingly used in K-12 classrooms, enabling teachers to facilitate more personalized learning experiences (S. Jiang et al., 2019; Kulik & Fletcher, 2016; Ma et al., 2014). The scalability of these platforms makes it easy to reach teachers and students across a broad range of educational contexts (Holstein et al., 2017; Purcell et al., 2013). However, these complex systems are rarely designed to be readily customizable by teachers, students, or other educational stakeholders (Drapeau et al., 2016; Kulkarni et al., 2014). With the diversity of textbooks, curricula, and teacher pedagogies, which frequently change over time, the instructional behavior of AI-based software can easily become misaligned with educational practice in a given context (Holstein et al., 2017). Thus, allowing teachers to customize and revise aspects of AI-based tutoring software may be essential for wider adoption and for sustaining their relevance in the long term (Aleven et al., 2016; Kulkarni et al., 2012; Matuk et al., 2016; Means et al., 2001; Nagashima et al., n.d.). For example, in a study that investigated the needs of K-12 teachers who use AI tutors in their classrooms, Prior work has found that teachers may discontinue the use of the AI tutors if there is a misalignment between instructional materials in the software and the curriculum in use, especially if there are no convenient ways for teachers to customize the content (Holstein et al., 2019b).

Crowdsourcing for education materials. Crowdsourcing can help create educational materials (Cross et al., 2014; Hills, 2015; Solemon et al., 2013), providing real-world educational experiences (Z. Chen & Luo, 2014; Dontcheva et al., 2014), exchanging complementary knowledge (Glassman et al., 2016), and providing feedback and evaluations for learners (Dow et al., 2013; Y. Jiang et al., 2018; Luther et al., 2015). For example, prior work used crowdsourcing to generate worked examples in mathematics (problems and step-by-step solutions) to volunteers using an open authoring tool (Aleahmad et al., 2009). They asked the volunteers to tailor the worked examples to the given student profiles that contained students' hobbies and skill proficiency. They found, interestingly, that math teachers wrote the best problem statements but that amateurs wrote the best solutions. Specifically for math hints, Prior work has crowdsourced math word problems and hints using a pipeline with built-in scaffolding of: 1) tagging components of crowdworkers' answers to a given template and 2) a faded, step-by-step tutorial (Y. Chen et al., 2016). Prior work has crowdsourced math tutorial videos that explain logarithms (Whitehill & Seltzer, 2017). They scaffolded crowds with examples of good video explanations, explicit guidelines, and contrasting examples of handwriting quality (given that the crowd may present their hand-written notes in the tutorial videos). Their crowdsourced tutorial videos lead to comparable learning gains as a popular Khan Academy video on logarithms (Whitehill & Seltzer, 2017). Similarly, trying to improve explanation in educational software, Williams et al. introduced AXIS that asks learners to generate, revise, and evaluate explanations as they solve a problem (Williams et al., 2016). Most of the studies seek to contribute new content, rather than to improve on or customize existing content (which can

be a higher bar for evaluation). The teacher-guided crowdsourcing pipeline (Chapter 3.1) involves teachers as expert facilitators to guide crowdwork, rather than merely as a comparison group as in (Aleahmad et al., 2009). Success stories from prior work bode well for open authoring and crowdsourcing of instructional materials. However, it remains an open question what level and kinds of expert (teacher) guidance can be effective in supporting crowd revision tasks. Prior work has not found a scalable, time-efficient way to enable stakeholders (teachers and students) to customize instructional materials in AI-based software (Drapeau et al., 2016; Glassman et al., 2016; Kulkarni et al., 2014). Work in Chapter 3.1 aims to address this open question through a crowd-based pipeline for instructional materials revision and customization.

Feasible Learning Group Formation Policies. Classroom customization involves not only tailoring instructional content and materials, but also designing the educational technologies that shape student interaction—for example, deciding which learning group formation (LGF) policies to implement in AI-based collaborative learning systems, which is the focus of Chapter 3.2.

In the computer-supported collaborative learning field, constructing effective, well-balanced learning groups is an important task (Amara et al., 2016; P. Dillenbourg, 2002; X. Wang et al., 2021). The importance of learning group formation (LGF) has been validated empirically (Huang et al., 2014; Webb et al., 1998). Literature on LGF in collaborative learning is vast. Researchers have paired students based on gender (Ding et al., 2011; Lockheed & Harris, 1984), learning style (Kuo et al., 2015; Sandmire & Boyce, 2004; Taniguchi et al., 2018), students' social network (Chuang et al., 2012), and their intelligence or task proficiency (Day et al., 2005; Niu et al., 2018; Sutherland & Topping, 1999). The majority of existing approaches to LGF do not support dynamic group formation (Amara et al., 2016). Dynamic group formation refers to the process of groups “created on demand while various domain-specific restrictions have to be considered” (Amara et al., 2016), or can “adapt to and benefit from previous information about group members and their abilities” (Maqtary et al., 2019; Mujkanovic et al., 2012). Compared to static, pre-planned LGF, the dynamic composition of groups allows for quick regrouping of learners based on up-to-date information regarding their progress and struggles. Dynamic group formation is an interesting issue, as researchers start envisioning more sophisticated and personalized classroom interactions (Holstein et al., 2019b) and more fluid social transitions (i.e., student social transitions that occur not all at the same time for everyone in the class) (J. K. Olsen et al., 2020), which are more challenging to orchestrate. Pairing policies that work well in practice need to be feasible, in the sense that enough partners can be found under the given grouping policies (i.e., good policy coverage). To this end, we explore a process we call “SimPairing”, to simulate pairing policies on authentic data and evaluate their feasibility (Chapter 3.2).

2.2. Human-AI Co-orchestration in the Classroom

Classroom orchestration broadly refers to systematic organization, continuous monitoring, and dynamic adjustment of instructional activities within learning environments (P. Dillenbourg & Jermann, 2010; Holstein et al., 2019b). Instead of focusing on interactions between an individual student or groups of students with an AI system (through the lens of AIED or CSCL), from the orchestration perspective, learner interactions are viewed at the level of a wider learning ecosystem, often comprising a classroom with one or more instructors, and varied groups of learners. Empirical studies suggest that the co-orchestration of classroom activities by human educators and AI systems could offer enhanced efficacy in pedagogical management (Sharples, 2013). Researchers have summarized five aspects of orchestration in their conceptual framework: adaptation, management, awareness, planning, and roles of the actors (Prieto et al., 2011), which connects to work on Pair-Up, where we leverage *adaptation* (on-the-spot tailoring of the plan or design during the learning activity) to provide dynamically differentiated collaborative learning and allow for *awareness* (understanding what is occurring in the learning environment) and *management* (continuous and in-the-moment regulation of learning activities) on the teachers' side.

Human-AI co-orchestration refers to distributing classroom management responsibility between (1) instructors and learners, (2) instructors and AI technology, or (3) instructors, learners, and AI (Holstein & Olsen, 2021). In co-orchestration, there are two models: 1) role splitting, where humans and AI take on different tasks; and 2) role sharing, where both contribute to the same instructional task (Holstein & Olsen, 2021). Some AI systems have been developed to “guide instructors’ attention to learners most in need of human attention while delegating support for other students to AI agents” (e.g., (Holstein, McLaren, et al., 2018; Martinez-Maldonado, Clayphan, Yacef, et al., 2015)). For example, the FACT orchestration real-time system alerts and recommends teachers about which students to help and what to help them with, as well as suggesting instructional responses that teachers can broadcast to a particular group or the entire class (VanLehn et al., 2018). Several systems, like FACT (VanLehn et al., 2018) and PyramidApp (Manathunga & Hernández-Leo, 2019), allow for some flexibility for students to work out of sync but mostly address the case of students who start the activity late rather than supporting asynchronous pacing by design (J. K. Olsen et al., 2020). One related work is by Molenaar et al., where they demonstrate a system for use in technology-enhanced classrooms that dynamically assigns students into different groups based on their current support needs to help them transition to self-regulated learning (Holstein & Olsen, 2021; Molenaar et al., 2019). With the relevant goal of exploring decisions made by different agents, Echeverria et al. conducted a technology probe study in the classroom for transitioning between individual and collaborative learning. They explored three pairing policies, respectively, with the pairing decided by the student, teacher, and AI

system (Echeverria et al., 2020). These shown, while AI-assisted classroom management is advantageous, it's often the preference of instructors and students to maintain a degree of autonomy and influence over classroom activities (Echeverria et al., 2020; Holstein, Aleven, et al., 2020; Yang, Lawrence, et al., 2021). My thesis work extends this thread of work in human-AI co-orchestration by studying a tool that helps teachers orchestrate dynamic combinations of individual and collaborative learning to achieve more **dynamically-differentiated** learning, in a way that respects teachers' preferences and boundaries (Chapters 4 - 6).

2.3 Dynamically Combining Individual and Collaborative Learning

Individual and collaborative learning modes may have complementary strengths for efficiently supporting learning (Echeverria et al., 2020; Mullins et al., 2011; Yang, Echeverria, et al., 2021a). For example, collaboration may increase students' learning outcomes by using instructional material that requires students to reason with mathematical concepts, as it promotes mutual elaboration (Diziol et al., 2007; Mullins et al., 2011). When learning mathematics, collaborative learning allows students to learn by asking for help and receiving explanations from a partner (Webb, 1989), which enables the student to fill knowledge gaps and correct misconceptions (Mullins et al., 2011). Knowledge co-construction has yielded improved student achievement in collaboration (Berg, 1994). However, the beneficial effects of collaboration on knowledge acquisition cannot always be found (Mullins et al., 2011; Souvignier & Kronenberger, 2007). Specifically, from a meta-analysis by Lou et al. (Lou et al., 1996), where they evaluated the impact of collaboration, about a fourth of the results showed no or even negative effects when compared to individual learning. Individual and collaborative learning may be beneficial under different math learning contexts. For example, in a study of grade-7 students solving algebra problems, results from two experiments indicated that individual learning was superior to collaborative learning when using worked examples (Retnowati et al., 2017). They've found that collaboration could be beneficial when learning under problem-solving conditions, but it may be counterproductive when studying worked examples (Retnowati et al., 2017). Given the complementary strengths, collaborative and individual learning are often combined in the classroom (e.g., Think-Pair-Share (Kothiyal et al., 2014) and Jigsaw (Aronson, 1978)). Research showed that combining individual and collaborative learning may be more effective than either mode alone, evidenced by students tending to make fewer errors and asking for fewer hints when engaged in a combination of individual and collaborative learning than in only collaborative or only individual learning (J. K. Olsen et al., 2019). It is important to understand when and how such a combination benefits students (J. K. Olsen et al., 2019); one possibility is dynamically combining them.

Students may work individually, in pairs, in large groups, or as a whole class; Olsen et al. refer to these different modes as different social levels in classrooms (J. K. Olsen et al., 2020). They argued that supporting fluid social transitions between the social levels benefits teachers, students, and the class as a whole (J. K. Olsen et al., 2020). Building on the idea of fluid social transitions (J. K. Olsen et al., 2020), I defined dynamically combining individual and collaborative learning as having at least three main features: 1) monitoring students in both individual and collaborative learning modes, 2) having students transition when the need arises (e.g., when students are no longer progressing productively in one mode of learning); and 3) pairing up students in ways that are not fully pre-planned. One potential benefit of such dynamic combinations is that they hold the potential to achieve **more personalized and differentiated learning**. Each learner is unique in their needs, abilities, learning paces, and ways of acquiring knowledge (Xie et al., 2019). It may yield better learning results if we engage learners in learning activities most suitable for them at the given moment. Managing such transitions is a challenging task for teachers, as they involve monitoring the real-time progress of students, determining the optimal transitioning time point and partners, and performing the transitions. Moreover, many tools today are specifically designed to support one social level (J. K. Olsen et al., 2020) by focusing exclusively on individual learning (Holstein, McLaren, et al., 2018), collaborative learning (Alavi & Dillenbourg, 2012; Cuendet et al., 2011; Mercier, 2016), or whole-class learning (Raca & Dillenbourg, 2013). Tools that do support multiple social levels have limited support for smooth transitions in between (J. K. Olsen et al., 2020) or provide support merely related to the timing of activities (Håklev et al., 2017; Martinez-Maldonado, Clayphan, & Kay, 2015; Martinez-Maldonado et al., 2013), still leaving the majority of the orchestration load for the teachers (J. K. Olsen et al., 2020). Specifically, the series of studies in Chapters 4 - 6 focuses on this topic in creating and evaluating tools that support such a dynamic combination of individual and collaborative learning, to achieve more personalized and **differentiated** learning environments.

2.4 MMLA used to support reflective teaching environments

Improving educational practice through reflection is an important focus of teacher professional development (Prieto et al., 2020). Prior literature review has found that teacher reflection is said to scaffold critical thinking, provide a source of knowledge construction in teaching, and promote self-regulation in teachers (Marcos et al., 2011). However, more work is needed to understand how a reflection tool can be designed to fit into teachers' routines and keep in mind teachers' practical need to maintain their workload manageable and to customize analytics (Prieto et al., 2020).

Prior researchers have studied how technology can aid teacher reflection. For instance, Romano and Schwartz studied teachers' preferred adoption and suggestions for three types of technology to support teacher reflection (Romano & Schwartz, 2005). They introduced

first-year teachers to three technology tools designed to elicit and encourage their reflections on teaching: (a) electronic portfolios, (b) online discussion, and (c) videotaping teaching. Results indicate that the first-year teachers in this study found value in each of the tools, with videotaping teaching encouraging the most meaningful reflection on their teaching practice.

MMLA state of the art. There is an emerging trend in AIED to collect rich, multimodal data sets in classrooms to inform the learning analytics tools. The advent of Multimodal Learning Analytics (MMLA) promises to enrich the understanding of educational contexts by integrating and analyzing data from both digital and physical learning environments, including physiological, behavioral, and environmental cues (Cukurova et al., 2020). This approach seeks to provide a more holistic view of the classroom (e.g., ClassInsight (Tamim et al., 2021)), capturing traces of teacher and student behaviors (such as student engagement, performance, cognition, and emotional states) that traditional digital analytics might overlook. Due to the extensive and complex multimodal data captured from these rich learning environments, AI models and data mining are becoming the most commonly used technologies to fuel MMLA applications (Blikstein & Worsley, 2016; Cukurova et al., 2019; Ouhaichi et al., 2023). For example, the cardiopulmonary resuscitation (CPR) tutor, which uses multimodal data streams (i.e., kinematic and electromyographic data) to predict mistakes and automatically provide audio feedback and support learners in improving their CPR performance (Di Mitri et al., 2020). TeamSlides, a team reflection tool, offers teachers MMLA visualizations to support reflective practices. Teachers valued its support in providing evidence while discussing teamwork and communication strategies (Echeverria et al., 2024). MMLA promises educators can gain insights into various factors affecting the learning process, including teacher actions and student engagement, thereby enabling a more informed approach to teaching and learning.

Implementing MMLA tools encounters challenges, including stakeholder preferences, privacy concerns, and the reluctance to share sensitive data (Alwahaby et al., 2022; Yan et al., 2022) School of Computer Sciences (Alwahaby et al., 2022; Yan et al., 2022). Cukurova et al. highlight that MMLA, venturing into complex territories with scarce guidelines and regulations, needs special consideration for vulnerable populations, such as children, so as not to contribute to a surveillance culture or raise privacy concerns (Cukurova et al., 2020; Ouhaichi et al., 2023). Research by Beardsley et al. introduced an informed consent comprehension test for MMLA studies, improving ethical practices (Beardsley et al., 2020). While analytics can offer valuable insights into teaching and learning processes, there is a significant gap in understanding which specific data elements and analytics are most beneficial for teachers (Borchers et al., 2023; Yan et al., 2022). **My research (Chapter 7) probes these challenges, seeking greater insight into stakeholders' preferences, acceptance, and privacy considerations regarding MMLA analytics. I do so in the context of developing and designing a teacher reflection tool.**

Given the nascent stage of MMLA applications, there have been few cases used in authentic classroom settings and they are mostly done in lab studies (Phillips et al., 2023). Recent literature reviews reported that MMLA solutions often focus on developing computational models, capabilities, and analytics for research purposes rather than on practical applications that close the analytic feedback loop (Alwahaby et al., 2022; Yan et al., 2022). More work is needed to understand how a reflection tool can be designed to fit into teachers' routines and keep in mind teachers' practical need to maintain their workload manageable and to customize analytics (Prieto et al., 2020). In both instances, involving teachers in the design process ensured the MMLA analytics were relevant for adoption into classroom practices. **My research in the Chapters 8 - 10 focused on: MMLA-based reflection tool (Chapter 8), and an exploratory study of it in authentic classroom settings (Chapters 9 - 10).**

I. CUSTOMIZATION

CHAPTER 3 CUSTOMIZED INSTRUCTIONAL MATERIALS AND TECHNOLOGY

This chapter describes the two studies focusing on the first of my three main themes, customization. These two studies broadly concern finding methods to make instructional materials (Study 1, Table 1.1) or technology (Study 2, Table 1.1) more customized to classroom contexts, including teachers' preferences.

Specifically, Study 1 aims to achieve customized instructional materials in AI-based tutoring software through a crowd-based pipeline and experiment. Study 2 aims to study how to fit and customize AI-based educational technology (specifically, dynamic pairing algorithms) to classroom contexts, and I achieve this through a simulation using historical, educational log data.

3.1 Pipeline Design for Teacher Customization of Materials

Acknowledgements. This work is under the mentorship and guidance of Prof. Kenneth Holstein, Prof. Vincent Aleven, Prof. Joseph Jay Williams. The project takes context of AI-based tutoring software, Lynnette, built by the CMU research team led by Prof. Vincent Aleven before I joined the team. The project received help (in coding data) from collaborator Tomohiro Nagashima and research intern Junhui Yao.

The work in section 3.1 is based on this CSCW publication¹. This project took inspiration from a prior study by other researchers that uncovered a critical teacher need: when using an AI tutor in classrooms, teachers often find that particular on-demand hints are insufficiently helpful for their students (Holstein et al., 2019b). In turn, teachers desire a simple way to revise these hints for the benefit of future students who reach a similar impasse. In co-design sessions with teachers, teachers envisioned mechanisms to provide rapid feedback to the AI tutor, even during an ongoing class session, as soon as a complaint about a particular hint arises, with minimal interruption of their teaching activity (Holstein et al., 2019b).

3.1.1 Research Questions

My work broadly concerns the following questions:

- RQ1. Can an expert-guided crowd pipeline lead to quality revisions that improve on the original artifacts, while saving experts' time?

¹ Kexin Bella Yang, Tomohiro Nagashima, Junhui Yao, Joseph Jay Williams, Kenneth Holstein, Vincent Aleven. Can crowds customize instructional materials with minimal expert guidance? Exploring teacher-guided crowdsourcing for improving hints in an AI-based tutor. In *Proceedings of the ACM on Human-Computer Interaction (CSCW' 21*, [PDE](#))

- RQ2. How does an expert-guided crowd revision condition compare to other comparison conditions?
- RQ3. What expert guidance is needed for crowdworkers to perform the revision tasks?
- RQ4. Can customization of instructional materials in AI-based educational software improve teachers' satisfaction?

3.1.2 Study Context: AI-based tutoring software

This study takes context in an AI-based tutoring system, Lynnette (Fig. 3.1.1(left)), which offers guided practice to middle-school students. I used the given ITSs as a platform for conducting studies. The scope of mathematical questions targeted in the experiment is middle school math algebra, specifically two-step equations (e.g., $2x + 4 = 0$, solve for x). It provides step-by-step guidance in the form of adaptive hints, correctness feedback, and error-specific messages. AI-tutors are increasingly being used in K-12 classrooms to help teachers more effectively personalize instruction (Ritter et al., 2007).

This workflow aims to improve on-demand hints in the software, as shown in Fig. 1. When using the AI tutor, students can request hints when stuck. These hints suggest the next steps and explain the underlying problem-solving principles, helping students understand key concepts. Authored by an expert with over 20 years of experience in intelligent tutoring systems (ITS), the hints are context-sensitive, concise, and emphasize algebra rules. The AI tutor (Lynnette) has been proven to improve students' equation-solving skills in several studies (Holstein, McLaren, et al., 2018; Long & Aleven, 2013; Waalkens et al., 2013). However, teachers often wish to edit and customize these hints, indicating unmet needs and opportunities for further improvement (Holstein et al., 2019a, 2019b).

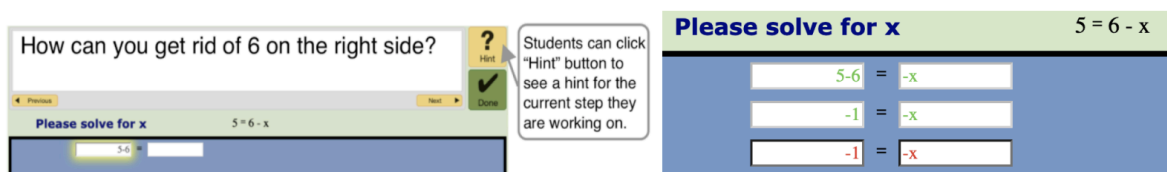


Figure 3.1.1 Left: An example hint level in the math AI tutor (Lynnette). Right: An example problem state.

I developed Timic, an expert-guided crowdsourcing pipeline for instructional materials improvement and customization (Figure 3.1.2). I next introduce the four stages in the pipeline.

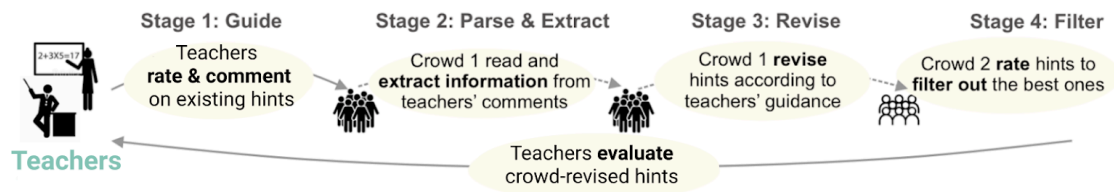


Figure 3.1.2 *Timic* - Teacher-guided crowdsourcing workflow for instructional materials (AI-tutor hints) revision

3.1.3 Pipeline Design

Stage 1: Guide. In Stage 1 (Figure 3.1.3, left), teachers acted as expert facilitators, providing ratings and comments on existing hints to guide crowd revisions. Teachers were familiarized with the AI tutor, shown how it works (Figure 3.1.1, left), and asked to try the system. They then rated and commented on five sets of AI-tutor hints, representing common critiques from previous studies (Holstein et al., 2019b). After this, teachers created their own preferred hint revisions and answered four open-ended questions on AI-tutor hint authoring.

Stage 2: Parse and Extract. In Stage 2, crowdworkers read teachers' guidance, including comments and ratings on hint effectiveness. Based on this, they listed changes to make and merits to keep in the existing hints (Figure 3.1.3, right).

Stage 1: Guide	Stage 2: Parse & Extract
For a typical student, how helpful do you think the hints above would be for students' learning? Not at all helpful 1 2 3 4 5 6 7 Extremely helpful (a) Expert Rate Imagine you are in class helping students and don't have much time... Please take 15 - 20 seconds to tell us: What would you like to see improved in these hints? (b) Expert Comment	According to teachers' comments, please use this space to list some: 1) Changes or improvement they want to see in the hints 2) Merits in existing hints. Changes or improvement Changes or improvement Changes or improvement Merit Merit (c) Crowd Parse

Figure 3.1.3. Left: Stage 1, Teachers Rate (a) and Comment (b) on Existing Hints; Right: Stage 2, Crowdworkers Read Teachers' Guidance and List the Desired Changes and Merits to Retain in the Hints to Be Revised

Stage 3: Revise. In Stage 3, conducted within the same online survey and HIT as Stage 2, MTurkers wrote or revised hints for five math problem states (Figure 3.1.4, left). Researchers approved HITs completed in good faith and rejected those without proper hints. Unqualified

hints included random inputs, copy-pasting, or mere answers without explanations (bottom-out answers were acceptable if accompanied by explanations).

The figure shows two side-by-side interfaces. The left interface, titled 'Stage 3: Revise', asks users to improve hints based on teachers' comments. It contains five text input fields labeled 'First hint level', 'Second hint level', 'Third hint level', 'Fourth hint level (Optional)', and 'Fifth hint level (Optional)'. The right interface, titled 'Stage 4: Filter', is a Likert scale evaluation form. It has a header with seven points: 'Extremely bad 1', 'Moderately bad 2', 'Slightly bad 3', 'Neither good nor bad 4', 'Slightly good 5', 'Moderately good 6', and 'Extremely good 7'. Below the header are three rows of questions, each with a radio button for each point: 'How mathematically correct/ accurate are the hints?', 'How easy to understand are the hints?', and 'Overall, how effective are the hints in helping students learn equation solving?'.

Figure 3.1.4. Left: Interface for crowd to write/revise hints; Right: Metric used for hint evaluation in 2 studies.

Stage 4: Filtering. For quality control, new MTurkers rated crowd-generated or revised hints on three dimensions: 1) mathematical correctness/accuracy, 2) ease of understanding, and 3) effectiveness in helping students learn (Figure 3.1.4, right). Inspired by prior crowdsourcing research, the question sequence began with objective questions to familiarize workers with the materials, potentially improving their performance on subjective questions. A seven-point Likert scale (1 to 7) was used. Workers justified their highest ratings. Each hint received evaluations from 8-12 crowdworkers, and researchers aggregated these ratings to select the highest-rated hints for expert evaluation.

3.1.3 Procedure

Participants. For Stage 1, I recruited 76 math teachers from an online forum. These teachers had varying levels of experience: 24 taught middle school, 7 taught high school, and 22 taught at multiple levels. Over 80% had more than 4 years of teaching experience, with most having 10 years or more. Teachers received a \$6 USD Amazon Gift Card for completing a 10-20 minute survey and were entered in a raffle for an additional \$40 USD. Valid responses were obtained from 53 teachers after excluding invalid ones.

For Stages 2 and 3, I recruited 41, and for Stage 4, 105 crowdworkers from Amazon Mechanical Turk, paying \$3 USD for a 15-25 minute survey. Participation was limited to US-based adult workers with a 98% or higher HIT approval rate and over 1000 approved HITs. I expected basic familiarity with middle school algebra.

For pipeline evaluation, 87 new math teachers from the same forum rated revised hints. Their demographics were similar to Stage 1 participants, with 27 teaching middle school, 22 high school, 5 higher education, and 33 at multiple levels. Over 80% had more than 4 years of teaching experience, with the majority having taught for 10 years or more.

Methods. Study 1 included 2 experimental and 3 comparison conditions:

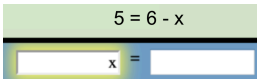
1. From-Scratch Generation (baseline): MTurkers wrote hints from scratch without guidance.
2. Rating Only Revision (experimental): MTurkers revised AI-tutor hints using teacher ratings.
3. Rating + Comment Revision (experimental): MTurkers revised AI-tutor hints using both teacher ratings and comments.
4. Randomly Selected Teacher-written Hints (comparison): Hints by randomly selected teacher experts.
5. Existing AI-tutor Hints (comparison): Original AI-tutor hints.

The study followed a four-stage workflow with the main manipulation in stages 2 and 3. A within-subjects design was used, where each crowdworker completed five problem states, assigned randomly to conditions 1, 2, or 3. This randomization ensured 12-14 crowdworkers per condition. In Stage 3, 105 workers rated all crowd-produced hints to filter the top-rated ones for expert evaluation.

Measures.

I measured the mean of teacher ratings on hint effectiveness to gauge teachers' satisfaction, a proxy for hint quality. Teachers rated five sets of hints for each of the five problem states in random order. Hints in conditions 1, 2, and 3 were filtered to the top 15 (5 problem states $\times$ 3 conditions). Teachers could also provide optional feedback. Table 3.1.1 shows a set of crowd-revised hints, expert feedback, and average teacher ratings.

Table 3.1.1: Example of highly-rated crowd-revised hints and expert feedback

Problem State and Existing AI-Tutor Hints	Example of one set of high-quality crowd-revised hints	Example teachers' comments on crowd-revised hints	Teacher rating
 Original hints: 1. You have constants on both sides: 5 on the left and 6 on the right. How can you have all constants on the left and no constants on the right? 2. How can you get rid of 6 on the right side? 3. Subtract 6 from both sides. 4. Write 5 - 6 on the left.	1. You have constants on both sides: 5 on the left and 6 on the right. We want to get all the constants on one side. How do we do this? 2. Let's get the constants together on one side of the equation: what should you do with the 6? 3. Let's get that X alone on the right side. Remember that whatever you do on the right side you also have to do on the left. (Rating + Comment Revision condition)	T13: "Again, I would have started by moving the variable to the left, but these are good, reasonable, clear hints" T52: "Third hint is perfect. This is how I state it in my class"	Mean: 5.77 out of 7

3.1.3 Results of Study 1 (Generic Hint Improvement)

To determine if providing guidance to crowdworkers required less time than teachers doing the revisions themselves (sub-question of RQ1), I compared the time taken for teachers to rate, comment on, and revise hints using log data from the survey platform. Averaging across the five problem states, it took teachers 37.2 seconds to read and rate a set of hints, an additional 57.6 seconds to comment, and an additional 145.6 seconds to rewrite them. This showed that teachers spent significantly less time guiding workers than revising hints themselves ($M = 37.2$ vs. 57.6 vs. 145.6 , $SD = 9.36$ vs. 9.86 vs. 23.22 , $df = 52$, $p < 0.001$). I performed a one-way ANOVA comparing teacher satisfaction with hints across five conditions, which revealed significant differences ($F(4, 1911) = 18.91$, $p < 0.001$). Post hoc paired t-tests indicated that crowdworkers' revised hints were better-rated than the original AI-tutor hints, although this improvement was not significant after Bonferroni correction (Table 3.1.2). To compare crowd-revised hints to those generated from scratch, I found that the revised hints were significantly better in the teachers' eyes (Table 3.1.2). Notably, no copy-paste behaviors were observed, and substantial revisions were made in each set. I compared crowd-revised hints to randomly selected teacher-written hints. Surprisingly, the crowd-revised hints in the Rating + Comment Revision condition were higher-rated than those written by randomly selected teachers ($M = 4.67$ vs. 5.00 , $SD = 1.83$ vs. 1.74 , $df = 86$, $p < 0.01$). The Rating Only Revision hints were also more highly rated, though the difference was not significant ($M = 4.67$ vs. 4.92 , $SD = 1.83$ vs. 1.64 , $df = 86$, $p = 0.06$). Finally, to investigate if expert comments facilitated crowdworkers beyond expert ratings alone, I compared teacher satisfaction with hints in the Rating Only and Rating + Comment Revision conditions and found no significant difference ($M = 4.92$ vs. 5.00 , $SD = 1.73$ vs. 1.64 , $df = 86$, $p = 0.48$).

In summary, the results (Table 3.1.2) demonstrated that it took teachers less time to provide minimal guidance to crowdworkers than to revise hints themselves. While this minimal guidance helped crowdworkers produce better hints than when generated from scratch, it was not sufficient to improve upon existing AI-tutor hints of relatively good quality. Additionally, the combination of expert ratings and comments did not significantly enhance the crowdworkers' performance over ratings alone.

Table 3.1.2. Mean and Standard Deviation (SD) for Each Hint Condition Evaluated in Study 1

	Con1: From-scratch Generation	Con2: Rating Only Crowd Revision	Con3: Rating + Comments Revision	Con4: Teacher-writt en hints	Con5: Existing AI-tutor Hints
Mean	4.02	4.92	5.00	4.67	4.79
SD	1.88	1.64	1.74	1.83	1.65

How much do teachers' preferences differ – and thus how great is the need for customization?

One factor that might limit the effectiveness of crowdsourced *generic* hint revision is possible variability in teachers' preferences. If teachers have different or conflicting preferences for hint content, then, given the generic approach to hint revision that was the subject of Study 1, a crowdworker might receive conflicting comments regarding how to improve the hints for a given problem state, leading to uncertainty as to where to go with the revision.² To determine if this is the case, I analyzed whether teachers demonstrate varying preferences to investigate if customization in AI-based educational software is needed (**RQ4**).

RQ4. Is customization of instructional materials in AI-based educational software needed to improve teachers' satisfaction? I started out by analyzing whether teachers' hint preferences converge or diverge, using their survey responses (n = 51) on four aspects regarding how math hints in an AI-tutor should be authored. The four aspects are 1) proper length of hints, 2) how conceptual or procedural hints should be, 3) whether hints should contain answers, and 4) whether hints should adapt to students' strategy. As shown in Figure 3.1.5 (upper part), teachers were relatively aligned in their preference for conceptual hints over procedural ones, and their preference for hints that adapt to student strategies versus not. They had more divergent preferences on the two other dimensions (hints' length, whether hints contain bottom-out answers) (Figure 3.1.5, lower part).

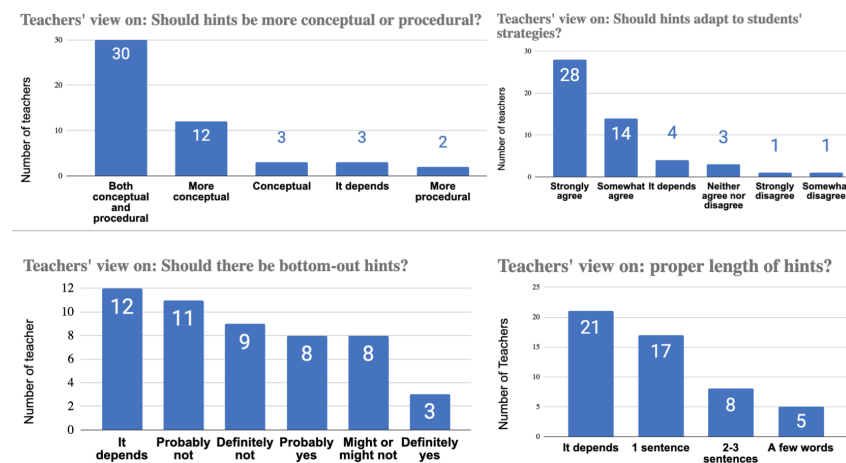


Figure 3.1.5. Teachers' Preferences on Hints: Preferences That Align (Upper) and Those That Differ (Lower)

Furthermore, from teachers' comments on existing AI-tutor hints, I found teachers not only have *different* views on how hints should be authored, but sometimes *conflicting* ones. For

² Due to randomization and platform limitation, I was unable to track and manually analyze which five comments each crowdworkers received.

example, while some teachers prefer hints to be *longer* because they can give more detailed explanations (e.g. T13, T29), some prefer them to be *shorter* as they think kids don't like to read much (e.g. T6, T11); while many teachers prefer hints to use mathematically *accurate terminologies* (e.g. T17, T22), many prefer *simpler* language (e.g. T9, T20); while many teachers think having answers in hints *don't* help kids develop transferable skills (e.g. T24, T47), some think it *does* (e.g. T23).

Additionally, teachers could have very idiosyncratic preferences that are hard to satisfy without customization. For example, they can be very specific about how long hints should be (“*The first hint should probably be 2-3 sentences at most. Any more than that & most students won't even bother to read it. The rest of the hints should be 1 sentence.*” (T17)), or how conceptual they should be (“*The first hint should be more conceptual, but the second hint should be more procedural but with some conceptual in it and the last hint pure procedural*” (T2)). These results show teachers can have different, highly specific and even conflicting preferences in math hints in AI-based educational software in some aspects, indicating teacher customization for instructional materials is important for satisfying the varying and nuanced preferences of end-users (teachers) in this context.

3.1.4 Results of Study 2 (Customized Hint Improvement)

Since I found teachers can have varied, highly specific, or even conflicting hint preferences, non-customized, generic fixes (i.e., fixes that try to address comments from multiple teachers at the same time) may not cater to everyone's taste. I conducted study 2 to investigate whether “*customized*” hints by crowdworkers (i.e., hints revised based on a single teacher's guidance) can lead to better teacher satisfaction. This experiment targets the following research questions:

RQ1. Can an expert-guided crowd revision pipeline improve upon existing AI-tutor hints, while saving experts' time?

In revisiting RQ1, I focus specifically on crowdsourced *customized* revisions to improve the existing AI-tutor hints.

RQ4. Is customization of instructional materials in AI-based educational software needed to improve teachers' satisfaction?

In revisiting RQ4, I focus specifically on teachers' *own* revised hints, compared to the existing AI-tutor hints.

Study 2 Results. Study 2 revealed that teachers found hints written by themselves to be better and more satisfying than existing AI-tutor hints. However, they did not perceive the crowd-produced customized hints as an improvement over the original hints. This suggests

room for improvement in AI-tutor hints and indicates that customization may increase satisfaction. However, crowd-customized hints, with minimal guidance, did not achieve this. Interestingly, teachers preferred their own written hints but still critiqued their edits made two months prior, with 45% containing some degree of critique.

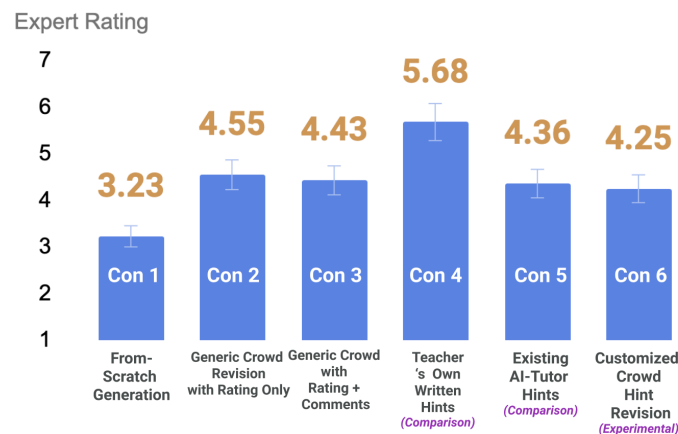


Figure 3.1.7. Teachers' Evaluation for Hints in Study 2

Teacher Complaints Unresolved by the Crowd: Complaints that persisted in both original and crowd-revised hints included the need for hints to ask conceptual questions and scaffold students (O1), mention the ultimate goal (O3), use accurate mathematical language (O4), and use familiar notations (O5). These issues often require domain knowledge and the ability to contextualize complaints, which the crowd may lack or fail to recall.

Teacher Complaints Resolved by the Crowd: Complaints successfully addressed by the crowd included removing bottom-out hints (O14). Crowdworkers fixed this by removing the last hint level or rewording to avoid giving direct answers.

New Issues Introduced by the Crowd: Newly introduced issues in crowd-revised hints included problems with clarity (O13), consistency (O8, O11), concreteness (O10), and mathematical accuracy (O12). These issues suggest areas where crowdworkers may have made the hints worse than the original ones.

Table 3.1.3: Experts' Complaints that Persist in Hints

Hint Source	Expert Complaints
Teachers think hints should	
Existing Tutor Hints	O1. Ask <i>conceptual questions</i> and <i>scaffold</i> students to reach answers on their own.
	O2. Explain <i>why</i> certain moves are appropriate.
	O3. Mention the <i>ultimate goal</i> and intent for solving this problem.

And Crowdsourced Revision	O4. Use more proper, mathematically <i>accurate language</i> .
	O5. Address the <i>highlighted step</i> students are currently on.
	O6. Use <i>notations</i> students and teachers are familiar with.
	O7. Use <i>simple language</i> to avoid overwhelming struggling students.

Table 3.1.4. Triangulation Results: Experts' Complaints Specific to Each Source of Hints

Hint Source		Experts' complaints
		Teachers do not like hints that
Crowdsourced Revision	Rating Only Revision	O8. <i>Skip steps</i> or stuff too many steps in one level because it might overwhelm students.
		O9. Seem hasty and <i>rushed</i> (in particular the later levels of the sequence)
	Rating + Comment Revision	O10. Stay at an abstract level. They want hints to be gradually specific and more <i>procedural</i> .
		O11. <i>Miss important information</i> which may lead to confusion or misconception.
	Customized Revision	O12. Contain typos, inaccurate expressions, or incorrect expressions.
		Teachers think hints should
Existing Hints in AI-tutor		O14. Delay giving answers until after more scaffolding, or not give answers at all.
		O15. Be more <i>flexible</i> and <i>contextualized</i> , and allow students to choose answer sequences, skip steps or use other valid strategies.

3.1.5 Findings from Two Experimental Studies

My studies evaluated an expert-guided crowd revision pipeline to increase teacher satisfaction with mathematical hints in an adaptive tutoring system, aiming to minimize teachers' time input. Key findings include:

1. **Time Efficiency:** Generating minimal guidance for crowds takes less time for teachers than revising hints themselves, suggesting minimally-guided crowdsourcing can save time.
2. **Hint Quality:** With minimal guidance, crowd-revised hints received higher teacher satisfaction compared to hints written from scratch by the crowd or randomly selected teachers. However, these revisions did not surpass the quality of the original hints, as rated by teachers.
3. **Teacher Preferences:** Teachers preferred their own revisions over those by the crowd or AI-tutor researchers, viewing only their own revisions as improvements. This suggests experts see room for improvement even in high-quality materials and highlights the importance of customization to individual preferences.

4. **Crowd Challenges:** Crowds struggled with the experts' specific requirements, often introducing new issues like clarity problems, inconsistencies, and inaccuracies. Effective crowd revisions require better scaffolding to prevent these issues.

Overall, the experiments show potential for saving teachers' time through crowd revision while emphasizing the need for customization to individual preferences and enhanced scaffolding for crowd workers. The goal is to integrate such pipelines economically into AI-based tutoring systems, benefiting educational technology companies by enabling software customization for diverse educational contexts.

Conclusion and How it Connects to the Thesis Theme

In sum, this work contributes two empirical studies that demonstrate the feasibility of a teacher-guided crowdsourcing pipeline for improving and customizing instructional materials, specifically, hints that can guide students' step-level problem-solving in an intelligent tutoring system. This pipeline aims to increase teacher satisfaction while minimizing their time input. This research addresses recent calls in CSCW, HCI, and educational technology for reporting null, negative, and positive results to avoid "file drawer" effects in the literature. There is a lack of adequate prior work in exploring the effectiveness of minimal expert guidance in supporting crowdworkers for educational content revision and customization. I analyzed two forms of minimal expert guidance and identified where additional scaffolding is needed. This study reveals that experts' preferences for hints in educational software vary significantly, indicating a potential to increase satisfaction through customization. However, designing expert-crowd revision pipelines faces several challenges: 1) capturing and communicating experts' preferences to crowdworkers, 2) helping the crowd learn relevant content and pedagogical knowledge, and 3) ensuring coherence and preventing new issues in revision tasks. While overcoming these challenges may not guarantee success, this work suggests they are critical enabling conditions.

This work contributes to the thesis theme of cultivating customized learning environments in AI-based tutoring software by exploring teacher-guided crowdsourcing for improving *instructional materials*. I focus on tailoring hints to individual teacher preferences, used within the context of tutoring systems. By using minimal guidance for crowds, I demonstrate significant time savings for teachers, allowing them to concentrate on more critical teaching tasks. This study addresses the diverse preferences of experts, highlighting the need for personalized and effective learning experiences. Additionally, I identify the necessity for better scaffolding to ensure the accuracy and relevance of crowd-sourced revisions. By providing empirical insights, I aid in developing robust customization processes for AI-based tutoring systems. Overall, this study supports the creation of a more personalized learning environment by improving the efficiency and quality of content customization in AI-based tutoring software.

3.2 Customizing Dynamic Group Formation Policy for Classroom Context

Acknowledgements. This study was done with collaborator Xuejian Wang, under the guidance of Dr. Vanessa Echeverria, Prof. LuEttaMae Lawrence, Prof. Kenneth Holstein, Prof. Nikol Rummel and Prof. Vincent Aleven. This project leveraged historical data from Datashop, collected by previous researchers who leveraged ITSs in classroom study. This study also received help from technical staff from the research team.

Classroom customization involves not only tailoring instructional content and materials, but also designing the educational technologies that shape student interaction—for example, deciding which learning group formation (LGF) policies to implement in AI-based collaborative learning systems, which is the focus for Chapter 3.2. Section 3.2 is based on Study 2 from Table 1.1, and this publication³.

Next, I describe a study on the customization of dynamic pairing algorithms through a simulation using historical educational log data. Specifically, the reason that this work relates to my thesis theme of creating *customized* learning environments is that this work stems from a need to investigate whether dynamic pairing policies can achieve a high enough paired-up rate to be feasible in the use of different classroom contexts. This investigation studies the different needs of different classroom contexts for four dynamic pairing policies, including how the parameter can be tuned, which may call for technology customization.

Although I am placing it under this chapter, this work also paves the way for later the theme of differentiation in my thesis, as the dynamic pairing policy I studied in this study, are designed with the intention of providing dynamically differentiated learning.

3.2.1 Study Motivation and Introduction

Constructing effective, well-balanced learning groups is an important task in computer-supported collaborative learning (CSCL) (Amara et al., 2016; P. A. Dillenbourg, 2002; X. Wang et al., 2021). The importance of learning group formation (LGF) has been validated empirically (Huang et al., 2014; Webb et al., 1998). For instance, prior work has found that group composition had a major impact on the quality of group discussion and students' test scores, both during group work and subsequent individual tests (Webb et al., 1998). The majority of existing approaches to LGF do not support *dynamic* group formation (Amara et al., 2016). Dynamic group formation refers to the process of groups “created on demand while various domain-specific restrictions have to be considered” (Srba & Bielikova,

³ Kexin Bella Yang, Vanessa Echeverria, Xuejian Wang, LuEttaMae Lawrence, Kenneth Holstein, Nikol Rummel, Vincent Aleven. (2021) Exploring Policies for Dynamically Teaming Up Students through Log Data Simulation. In *Proceedings of the 14th International Conference on Educational Data Mining*. (EDM' 21, [PDE](#))

2015), or can “adapt to and benefit from previous information about group members and their abilities” (Maqtary et al., 2019; Mujkanovic et al., 2012). Compared to static, pre-planned LGF, the dynamic composition of groups allows for quick regrouping of learners based on up-to-date information regarding their progress and struggles. Dynamic group formation is an interesting issue, as researchers start envisioning more sophisticated and personalized classroom interactions (Holstein et al., 2019b) and more fluid social transitions (i.e., student social transitions that occur not all at the same time for everyone in the class) (J. K. Olsen et al., 2020), which are more challenging to orchestrate.

In the context of an Intelligent Tutoring System (ITS) that supports both individual and collaborative learning, it is useful to investigate whether dynamically switching students between the two modes, as the need arises, can be effective and feasible. Pairing policies that work well in practice ideally have characteristics of both effectiveness and feasibility. By effective, I mean that the pairing policy leads to students’ reaching desired learning goals, and by feasible, I mean that enough partners can be found under the given grouping policies (i.e., good policy coverage). Specifically, I defined *feasibility* as the percentage of students who can be teamed up under a given pairing policy.

The feasibility of LGF is an important issue to investigate in designing orchestration tools for teachers, and can be a central concern at the initial stage of tool design. This is because during the initial design stages, I often do not yet have data to rigorously evaluate the effectiveness of LGF, given that testing the LGF requires human resources of learners, instructors, materials resources of devices, systems, and a long time period. An effective pairing policy that only covers a small percentage of students in a classroom may have limited influence on the whole class. Thus, the feasibility of LGF can be important in providing context for the potential coverage of LGF in a class.

Literature on LGF in collaborative learning is vast. Researchers have paired students based on gender (Ding et al., 2011; Lockheed & Harris, 1984), learning style (Alfonseca et al., 2006; Kuo et al., 2015; Sandmire & Boyce, 2004; Taniguchi et al., 2018), students’ social network (Chuang et al., 2012), and their intelligence or task proficiency (Day et al., 2005; Niu et al., 2018; Sutherland & Topping, 1999). Heterogeneous and homogeneous group formation are two main approaches in team formation, and many studies have demonstrated their effectiveness in CSCL (Amara et al., 2016; Day et al., 2005; Fuchs et al., 1998; Maqtary et al., 2019; Storch & Aldosari, 2013; Sutherland & Topping, 1999). Students’ knowledge is argued to be the most suitable and important attribute to form educational groups (Maqtary et al., 2019). Prior work has also used machine learning or other algorithms to incorporate multiple factors for optimizing team formation (B. Chen et al., 2020; Y.-T. Lin et al., 2010; Pang et al., 2014; Tien et al., 2013). However, the literature on LGF provides little insight into the feasibility of these LGF policies, especially in the ITS context.

Evaluating the feasibility of LGF policies offline, before implementing them in real classrooms, is challenging, given a lack of readily accessible approaches. To address this problem, I adopt a process I call “*SimPairing*” to simulate pairing policies on authentic data and evaluate their feasibility. In this process, I used transaction data from several classes of students using an ITS, collected from classroom studies conducted in U.S. middle schools. I computed and analyzed how the feasibility of several LGF policies (described below) changed as each class progressed, and how the feasibility varied across different classes. Replaying historical data to simulate possible futures (e.g., Replay Enactment (Holstein, Harpstead, et al., 2020)) has been used as a method by researchers to design tools with similar data-driven, human-centered approaches (Nagashima et al., n.d.). Diana et al. (Diana et al., 2017), for instance, used machine learning (ridge regression) to predict students’ grades based on historical data in CS education (i.e., programming). Based on these *predicted grades* and simulated *students’ “helped” status*, they determined which students needed help and could provide assistance. They then used a network graph of code-state to search for potential peer tutors who shared a common ancestor node with the tutee. They found that grouping low-performing students and using better model features can increase the number of students helped. Their findings suggest that using low-level log data to group and match low-performing students with a peer tutor may effectively increase the amount of help given in a classroom. In contrast, in a mathematics education context, I simulated different policies selected based on literature and teachers’ common practice revealed in user-centered research with K-12 teachers (Echeverria et al., 2020; Olsen et al., 2020; Yang et al., 2021).

One novelty of the current work is that it studies *dynamic pairing policies that consider students’ in-the-moment wheel-spinning status*. Identifying students who are unproductively struggling, yet failing to master the skill (i.e., wheel spinning), is a first step to getting them unstuck (Beck & Gong, 2013). While there has been significant work on modeling and predicting wheel spinning (Kai et al., 2018; Matsuda et al., 2016; Zhang et al., 2019), little work has been dedicated to developing interventions to get them unstuck, with a few recent exceptions (Holstein et al., 2019a; Mu et al., n.d.). While a typical classroom has students struggling with problems and those who have excelled on the same problem, the latter students’ expertise is rarely utilized. Instead, often the only source of help is the instructor, who is likely unable to help all the students who need help within the time constraints of the class period (Diana et al., 2018). Peer tutoring (i.e., pairing a struggling student with a peer tutor) could be an effective way to help get struggling students unstuck when the instructor has their hands full.

Lastly, instead of prescribing *a specific grouping criterion*, my work envisions that instructors will customize pairing policies and parameters to their classroom contexts, which prior work argued to be especially helpful in the LGF process (Amara et al., 2016; Echeverria et al., 2020; Maqtary et al., 2019; Yang, Nagashima, et al., 2021). Amara et al. found that most of

the proposed LGF solutions do not allow instructors to customize the grouping process (Amara et al., 2016). They argued that it is less helpful to apply a grouping solution for all types of learners and more useful to leave the choice to instructors. Instructors can then form groups according to different learning objectives, learners' needs, activity types, and customize the LGF process according to location and time (Amara et al., 2016). Similarly, Echeverria et al. envision adaptability in an orchestration system, which "enables teachers to select the best pairing policies based on their particular goals, needs, and classroom dynamics" (Echeverria et al., 2020), to be helpful for different classrooms. In the current investigation, three of the four policies I studied involve an adjustable pairing threshold or parameter, which I simulated with various values.

In sum, the current work investigates the feasibility of four dynamic LGF policies derived from user research with math teachers. I investigate from three angles: overall session simulation, class-level variance, and session-level contrasting cases. This work contributes to the feasibility results of the dynamic pairing policies and recommendations for orchestration tool design, and it highlights future work regarding tools supporting dynamic LGF.

3.2.2 Methods

Based on simulations with historical transaction data from Lynnette, I evaluated how feasible four pairing policies (described below) are. I applied each pairing policy to data from each class session. For every minute in a session, I calculated the percentage of students who met the policy's criterion for being teamed up. Based on this calculation, I evaluated policy feasibility using two measures, FI_1 and FI_2 , defined in 3.2. In the simulation process, I did not make assumptions about how long simulated collaboration episodes would last. I foresee that in any of these episodes, students will be tasked with collaboratively solving several math problems; it is hard to predict how long that will take them. I thus did not simulate taking tutors or tutees out of the pool of students available for teaming up, or returning them to this pool, at the beginning and end of collaborative episodes, respectively. Although this simplification might introduce some inaccuracy into the simulation results, it may be hard to do better. As well, the asymmetric roles that paired-up students have in the pairing policies may limit the inaccuracy. For example, simultaneously keeping a struggling student and a non-struggling student in the pool instead of taking them both out might have offsetting effects in terms of feasibility.

My simulation involved four pairing policies, namely:

Policy 1 - Struggle with Non-Struggle: Pairing students who are wheel-spinning (unproductive struggle) with students who are not wheel-spinning.

Policy 2 - Pairing with Restriction: Pairing students who are wheel-spinning with those who are not, with a varying pairing restriction (PR) rate β . The PR rate simulates restrictions regarding who can collaborate with whom, which the teacher would provide in real life.

Policy 3 - Knowledge Difference Pairing: Pairing students whose knowledge levels (as measured by the tutor's BKT) differ by *more* than a certain threshold α .

Policy 4 - Knowledge Similarity Pairing: Pairing students whose knowledge levels (as measured by the tutor's BKT) differ by *less* than a certain ceiling γ .

The distinction in these policies aligns with Amara et al.'s categorization for dynamic group formation (Amara et al., 2016): *intra-session* and *inter-session* grouping. Intra-session grouping allows for changing group members during the learning process, which is useful, for example, for synchronous mobile collaborative learning (Amara et al., 2016). In *inter-session grouping*, groups are formed only before starting or after ending the learning process. Specifically, policies 1 and 2 fall under *intra-session* grouping since I simulated pairing students based on their in-the-moment struggle. These two policies also concern *fluid social transitions* (J. K. Olsen et al., 2020), since the students in a given class may transition from individual to collaborative learning at different times. My pairing policies 3 and 4 concern inter-session grouping and pair students based on their initial knowledge level. To apply these policies, teachers or the tutoring system would assess students' knowledge level before (or at the beginning of) a class session.

The research questions I aim to answer are:

RQ1: Based on a pairing simulation done with students' historical transaction data, how feasible are the four pairing policies?

RQ2: How does varying the parameters in the pairing policies affect the feasibility of pairing students?

RQ3: Does the feasibility of the pairing policies vary for different classes or sessions? If so, how?

3.2.3 The Four Pairing Policies

Policy 1: Struggle with Non-Struggle

Description. Policy 1 utilizes the struggle detector to pair wheel-spinning students with those who are not. The struggle detector assumes students' wheel-spinning status as a binary value for a given timestamp. Inspired by the work of Diana et al. (Diana et al., 2017), I categorized students in the *Struggle Pool* if they were wheel-spinning on at least one KC, indicating they

could need help from a partner. Students not wheel-spinning on *any* KC were categorized in the *Tutor Pool* and considered as available tutors. I simulated pairing students in the *Struggle Pool* with students in the *Tutor Pool*. To determine the feasibility of this policy, I calculated the percentage of struggling students who had a potential partner (for more details, see below).

Rationale. Literature suggests that when students are wheel-spinning, giving them more of the same type of math problems to solve may not be productive (Mu et al., n.d.). When wheel-spinning, students would likely benefit from instructor attention or extra instruction. However, prior user research in the classroom (e.g., (Holstein et al., 2019b)) found that teachers often cannot help all struggling students. In this case, wheel-spinning students may benefit from a peer tutor's help, which leads to a policy that seeks to dynamically find them partners (Mu et al., n.d.).

Policy 2: Pairing with Restriction

Description. Policy 2 is an extension to Policy 1, where I pair a struggling student with a non-struggling student, while enforcing a constraint that not all students are eligible for teaming up. The proportion of ineligible students is captured as the Pairing Restriction (PR) rate. The PR rate is used to simulate situations where the teacher prefers that certain students do not work together. Specifically, I simulate pairing students in the *Struggle Pool* with students in the *Tutor Pool*, while enforcing the restriction that $\beta\%$ ($0 < \beta < 1$, step = 0.1) of students in the *Tutor Pool* are ineligible as partners. For example, a PR rate β of 0.2 means 20% of the students in *Tutor Pool* have been restricted from working with any students in *Struggle Pool*. It is important to know how these restrictions affect the feasibility of the policies.

Rationale. I designed this policy based on results found in a survey I conducted with 54 middle-school math teachers on their pairing preferences in collaborative learning (Yang, Lawrence, et al., 2021) and semi-structured interviews conducted with middle school teachers. Teachers expressed a desire to set constraints so that certain pairs of students are restricted from working together. Previous studies and user research by Olsen et al. and Echeverria et al. also informed the idea of ruling out certain pairings in advance (Echeverria et al., 2020; J. K. Olsen et al., 2020). Such restrictions usually arise from information or concerns teachers have about their students' traits, behaviors, and interpersonal relationships (Maqtary et al., 2019).

Policy 3: Knowledge Difference Pairing

Description. In Policy 3, I pair students who have *different* Initial Knowledge (IK) levels. In a practical scenario, teachers may assess students' knowledge through quizzes or exams.

Alternatively, if the classrooms use ITS, teachers may have students practice several math questions individually, prior to transitioning into collaborative learning activities.

To simulate this policy without having pre-assessment data, I used data from the tutoring sessions (captured in the log data) to compute students' IK levels. Specifically, I computed a student's IK for each KC as the average mastery for their first *three* opportunities for this KC. The reason is that I want to use up only a small portion of the data from the tutoring session, so the measure represents initial knowledge. In these datasets, three opportunities generally fall in the first quartile (25%) of students' opportunities for any given KC. Another reason I chose the cutoff of three is that a previous EDM study with ASSISTments data showed student learning often appeared to occur after students had had *ten* opportunities with the target knowledge (Almeda, 2018). Thus, one may assume learners to have little learning on their first three times in transaction data, practicing a KC. A student's overall IK $S_j (j \in N)$ is calculated as the average of their IK across KCs. To more accurately calculate students' IK, I limit the simulation to sessions that practiced the first (i.e., the most basic) level of KCs, involving 25 sessions.

KD was the difference between the two students' IK, and was denoted as $S_{jk} (j, k \in N)$, which was calculated as the absolute value of the differences between the two students' IK:

$$KD(S_{jk}) = |IK(S_j) - IK(S_k)|$$

Inspired by prior work, which proposed a clustering LGF method that considers a threshold of learner heterogeneity (Huang & Wu, 2011), this work similarly considers a KD threshold. For this policy, the required KD of two students (S_1, S_2) should be a *minimum* of α ($0 < \alpha < 1$, step = 0.1) for them to be eligible to pair up.

Rationale. Findings from user research with math teachers informed the heterogeneous pairing policy. In the survey conducted with 54 math teachers, I found the most common way teachers paired students was pairing those with a different knowledge level (67%, $N = 34$) (Yang et al., n.d.). In this study, I use students' *mastery of knowledge components* (i.e., targeted math skills) calculated based on the BKT model to represent students' knowledge. In a systematic literature review on LGF in CSCL, Maqtary et al. found that the knowledge level is the most commonly used attribute in LGF, which they claim to be the most suitable and important attribute to form educational groups because of its effects on the group process (Maqtary et al., 2019).

There is a range of research that shows heterogeneous grouping can promote positive interdependence, better group performance, and effective interactions (Amara et al., 2016;

Huang & Wu, 2011; Hübscher, 2010; Johnson & Johnson, 1987; Webb & Palincsar, 1996). Heterogeneous group composition not only enhances elaborative thinking but also leads learners to deeper understanding, better reasoning abilities, and accuracy in long-term retention (Huang & Wu, 2011; Johnson & Johnson, 1987). Research also suggests that collaborative learning with heterogeneous group composition by characteristics such as gender, ability, achievement, social-economic status (SES), or race, can be beneficial (Webb & Palincsar, 1996).

Policy 4: Knowledge Similarity Pairing

Description. Policy 4 is analogous to Policy 3, with the same definition of KD and IK. To pair students with *similar* knowledge, using the same calculation as Policy 3, this policy simulated pairing students who have *a small* KD. To be eligible for students to form a pair under this policy, the KD of two students (S1, S2) should be *less than or equal to* γ ($0 < \gamma < 1$, step = 0.1). For example, when $\gamma = 0.2$, two students with knowledge of 0.6 and 0.75 (KD = 0.15, below γ) would be eligible to pair, but another pair with knowledge of 0.5 and 0.8 (KD = 0.3, above γ) would not be eligible.

Rationale. Policy 4 was inspired by prior literature and informed by user research. Literature suggests that homogenous groups can be beneficial for students' learning. For example, Fuchs et al. found that homogenous dyads generated greater cognitive conflict and produced better quality work than heterogeneous groups (Fuchs et al., 1998). Additionally, among 54 teachers I surveyed, 43% reported that they pair students with a similar level of knowledge (Yang et al., n.d.). This was the third most popular grouping method that teachers commonly adopt (43%, $N = 23$), following strategies of pairing students with different knowledge (Policy 3) and pairing students randomly (Yang, Lawrence, et al., 2021).

3.2.4 Metrics

In this section, I describe the metrics to evaluate the pairing policies. I discuss how prior work informed the metric definitions and how different metrics could be suitable to evaluate different policies. I build on Diana et al.'s work (Diana et al., 2017), who defined an Efficiency Index (EI) as a measure of a pairing algorithm's performance, specifically:

$$EI = \frac{LowPerformingStudentsHelped/BeingHelped}{LowPerformingStudents}$$

I adapted EI into two metrics of interest for the pairing policies: Feasibility Index 1 and 2. FI_1 is the percentage of students who can be paired among all struggling students in a session.

$$Feasibility\ Index - 1\ (FI_1) = \frac{StrugglingStudentsCouldBeHelped}{TotalStrugglingStudents}$$

FI_2 is the ratio of paired students among all the students in a session.

$$\text{Feasibility Index - 2 } (FI_2) = \frac{\text{StudentsPaired}}{\text{TotalStudents}}$$

For Policies 1 and 2: Given the goal of pairing all struggling students in the session, FI_1 was a suitable measure for policy feasibility, showing what percentage of students who are wheel-spinning can get help. *For Policies 3 and 4:* Given the goal of pairing all students in the session who satisfied a certain KD, FI_2 was a suitable measure for policy feasibility, as it calculated the percentage of the paired students out of the total students.

3.2.5 Class-Level Variance Analysis

I explored how the four pairing policies worked for different classes and whether they should be adapted to class-level differences (RQ3).

Class-level Differences: Building on Echeverria et al.'s insight that pairing support should adapt to different classroom contexts (Echeverria et al., 2020), I analyzed if there were systematic differences between classrooms and how these related to policy feasibility. The key context variables were students' struggle status (Policies 1 and 2) and initial knowledge (Policies 3 and 4).

Student Struggle Status: I calculated the number of struggling students (wheel-spinning) per minute in each class session, then computed Average Number of Struggling Students (ANSS) and the Average Struggling Students Ratio (ASSR) (Figure 3.2.1). One-way ANOVAs revealed significant differences in ANSS among classes ($F(5,62) = 4.34$, $p < 0.001$), with post hoc Tukey tests showing significant differences between C3 and C1 (diff = 2.55, $p < 0.001$). A marginally significant difference was found in ASSR between C3 and C2 (diff = 0.10, $p = 0.08$), indicating class-level differences in student struggle status.

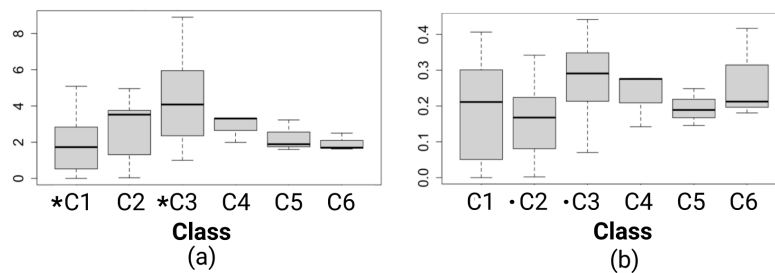


Figure 3.2.1 ANSS (a) and ASSRs (b) for Six Classes

Initial Knowledge: I measured each student's initial knowledge (IK) for all KCs involved, finding significant class effects on IK ($F(5,320) = 5.895$, $p < 0.05$). Post hoc Tukey tests showed significant differences between C2 and C1 (diff = 0.12, $p < 0.05$), C3 and C2 (diff =

-0.095, $p < 0.05$), and C4 and C2 (diff = -0.189, $p < 0.05$). C2 had the highest median IK (Figure 3.2.2).

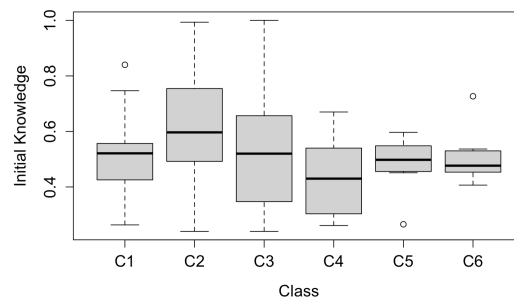


Figure 3.2.2. Initial Knowledge for Six Classes

Policy Feasibility for Policies 1 and 2: Policy 1 maintained a consistent high feasibility (FI1 above 0.85) across classes, with no significant class effect ($F(5,62) = 1.24$, $p = 0.30$). This suggests Policy 1 does not need class-level adaptation. Policy 2 showed varying feasibility with increasing PR rates across classes. ANCOVA results indicated significant differences in FI1 for eight class pairs ($p < 0.05$), suggesting Policy 2 may benefit from adaptation to class-specific characteristics, particularly regarding student struggle status (Figure 3.2.3).

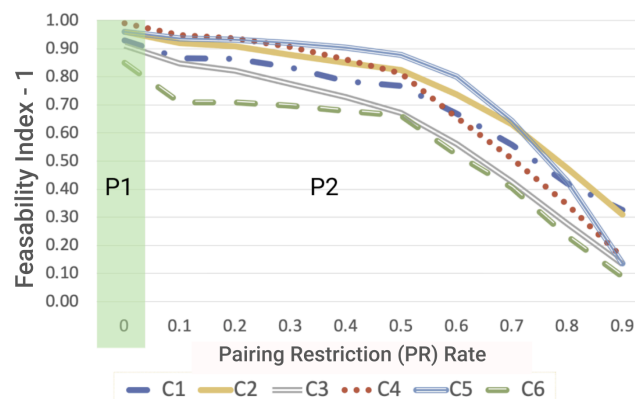


Figure 3.2.3. FI₁ of Policies 1 and 2 for Six Classes

Policy Feasibility for Policies 3 and 4: For Policy 3, FI2 decreased with different slopes for each class as knowledge heterogeneity increased. For example, at a KD of 0.1, class 6's FI2 dropped to 50%, while other classes remained above 75%, indicating different impacts of knowledge heterogeneity on policy feasibility (Figure 3.2.4).

To test whether Policy 3 behaved differently for each class, I conducted ANCOVAs with Class as a categorical variable, KD as a quantitative variable, and FI2 as the outcome

variable. Comparing models with and without a Class $\times$ KD interaction term showed no significant interaction effect ($F = 0.81, p = 0.54$). Using an additive model, I found significant differences in FI2 for three class pairs (C1-C3, C3-C5, C3-C6) with $p < 0.05$, and marginal differences for two pairs (C2-C5, C2-C6) with $p < 0.1$.

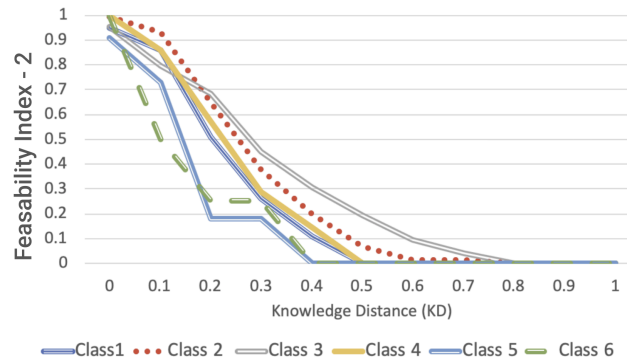


Figure 3.2.4. FI₂ of Pairing Policy 3 for Six Different Classes

For Policy 4, the six classes were more convergent, indicating less class-level difference than Policy 3. ANCOVA tests confirmed this. Comparing models with and without a Class $\times$ KD interaction term showed no significant interaction effect ($F = 0.13, p = 0.99$). Using an additive model, there were no significant differences in FI2 ($p > 0.05$) for Policy 4, suggesting Policy 4 performed consistently across classes and does not need adaptation (Figure 3.4.5).

I also examined the relationship between feasibility variance for Policies 3 and 4 and class-level IK characteristics. Significant IK differences were found between C2-C1, C3-C2, and C4-C2. However, these differences in IK could not predict feasibility differences for Policies 3 and 4 accurately, indicating that other classroom characteristics might be needed to represent class variance in policy feasibility.

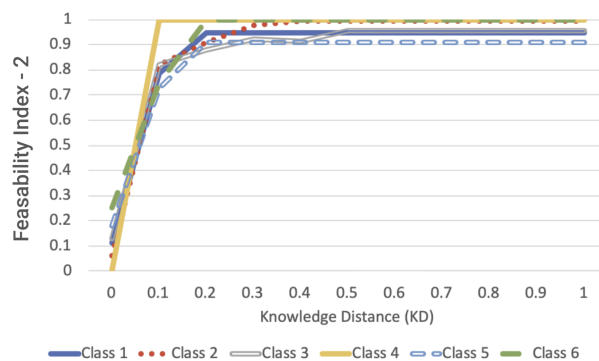


Figure 3.2.5. FI₂ of Policy 4 for Six Different Classes

3.2.6 Analysis of Contrasting Cases

I conducted a case study to understand how policies may perform dynamically (e.g., across every minute during class time) and differently in different class sessions (RQ3). For every policy, I selected a *typical* case and an *extreme* case in terms of the policy feasibility simulation results. For the *typical* case for all four policies, I selected a session (Session 1, C1) that had an average length of time (i.e., 41 minutes) and an average number of students (i.e., 13 students). In the session, policies are typically performed (as shown by visually comparing the simulation results of each policy for all sessions). As for the *extreme case*, I examined the simulation results for each policy on each session and identified different sessions where each policy performed surprisingly or differently from the common trend. The extreme case can be a worst-case scenario (Policies 1, 2, and 3) or one that works surprisingly well (Policy 4). Below, I present the analysis and results for these contrasting cases for each policy.

Policy 1. In Policy 1, I chose the extreme case (Session 19, C3) as it was a session on which this policy had the worst performance, and thus it had the most different FI_1 trend, from examining visualizations of FI_1 for all sessions involved. I compare the typical case and extreme case by first contextualizing the struggle status of the two cases, and comparing the visualization of feasibility (for each minute) in the two sessions. Figure 3.2.6 depicts the *ratio of struggling students* (among all students in the class session) for the contrasting cases. For Policy 1 simulation (Figure 3.2.7), I obtained, for every minute in the class session, three values regarding policy feasibility: the number of students who were not wheel-spinning on any KCs (green bar), the number of students who were struggling, and *had* a potential partner (yellow bar), and the number of students who were struggling and *did not have* a potential partner (red bar).

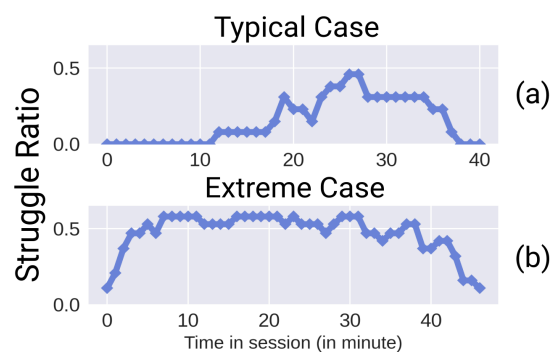


Figure 3.2.6. Struggle Ratio of a Typical (a) and Extreme (b) Case

Typical Case. In the typical case, students started to struggle after 10 minutes, as shown in Figure 3.2.6 (a) and Figure 3.2.7 (a). In all instances of wheel-spinning, a potential partner was available (i.e., $FI_1 = 1$ for any minute in this session). The typical case aligned with the

overall simulation results from Policy 1, which showed that, for an average class, most struggling students could find potential partners the minute they struggled.

Extreme Case. Among all sessions in the dataset, the per-minute struggle ratio rarely goes over 50%. By contrast, the extreme case session had more struggling students than non-struggling students in 27 out of 46 minutes, indicated by a struggle ratio of above 0.5, as shown in Figure 3.2.6 (b). This resulted in lower feasibility for Policy 1. The extreme case differs from the typical case in two aspects. First, unlike the typical case, almost as soon as the class began, students started wheel-spinning. Second, there were wheel-spinning students without potential partners in almost every minute of the session (indicated by red bars in Figure 3.2.7 (b)).

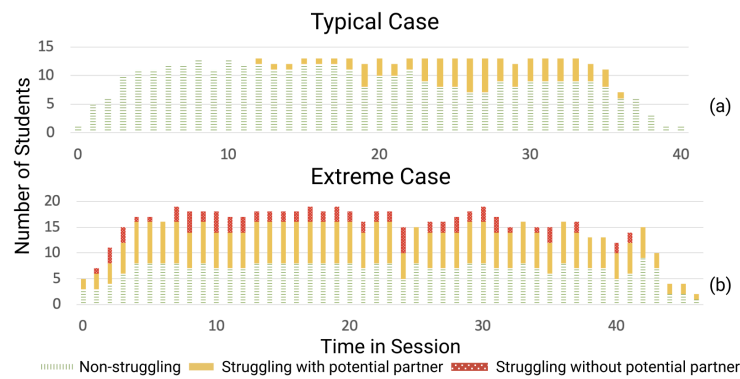


Figure 3.2.7. Policy 1 for a Typical (a) and an Extreme Case (b)

Policy 2. As with Policy 1, I chose this extreme case (Session 19, C3) as this policy had the worst performance in this session. This session also had the most different FI_1 trends. I simulated Policy 2 and calculated FI_1 for every minute in the two contrasting class sessions. Figure 3.2.8 shows the typical and extreme cases of how FI_1 changed when different PR rates were simulated. I plotted four different PR rates in the figure.

Typical Case. I saw two patterns in the Policy 2 simulation for the typical case. Firstly, the policy was typically robust in maintaining high feasibility with a non-zero (albeit low) PR rate. In Figure 3.2.8 (a), lines with PR rate 0.1 and PR rate zero completely overlapped. With these PR rates, there were no instances of struggle without a potential partner (i.e., feasibility was 1 across the whole session). Secondly, when the PR rate was high (0.5 or 0.8), FI_1 exhibited a sharp decrease when there was an increase in student struggle. For instance, in Figure 3.2.6 (a) at minute 19, the struggle ratio increased from 0.07 to 0.23, as the number of wheel-spinning students went from 1 to 3. In Figure 3.2.8 (a), at the same time ($t = 19$ min), I saw a sharp decrease in FI_1 when the PR rate was 0.8.

Extreme Case. As shown in Figure 3.2.8 (b), the extreme case exhibited very different patterns compared to the typical case, mainly in three aspects. First, given that it had a higher struggle ratio, even when there was no pairing restriction (i.e., PR rate = 0), I observed that the FI_1 was not always 1 or even close to 1, as I saw in the typical case. Second, even a slight PR rate of 0.1 further worsened the policy feasibility and lowered the FI_1 , unlike the typical case, which showed resistance to a low PR rate. Third, if a class had a higher struggle ratio, the PR rate had a stronger effect on worsening FI than for a session that had a lower struggle ratio. This effect was especially prominent when the PR rate was high (e.g., 0.5 or 0.8). This contrast means that the instructors may afford to set a higher PR rate without affecting the FI_1 too much, for a common session that has a moderate struggle ratio. However, the instructors may need to consider lowering the PR rate for a high-struggle session.

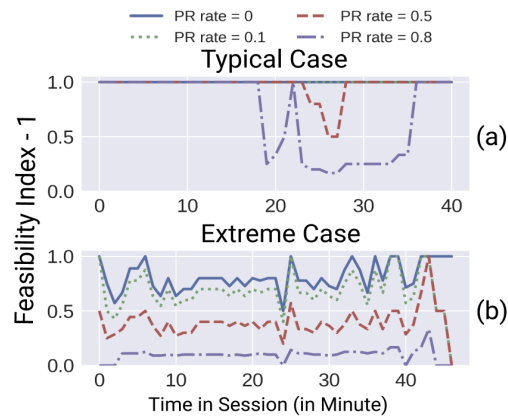


Figure 3.2.8. Policy 2 for a Typical (a) and Extreme (b) Case

Policy 3. In Figure 3.2.9, I present the results for Policy 3 simulation on two contrasting cases, plotting FI_2 for every step of the knowledge distance threshold for that session. The extreme case was chosen for having the most different FI_2 trend, from examining the visualizations of FI_2 for all sessions involved.

Typical Case. As shown in Figure 3.2.9 (a), for the typical case, the FI_2 dropped gradually as the required KD threshold increased, which aligned with the overall simulation result. To pair students based on different knowledge (Policy 3), the instructors need to balance the required heterogeneity (i.e., higher knowledge distance threshold) and the desired paired ratio of the whole class. In this typical case, if a teacher selects a threshold of 0.5 or higher, none (0%) of students in the class session would be paired.

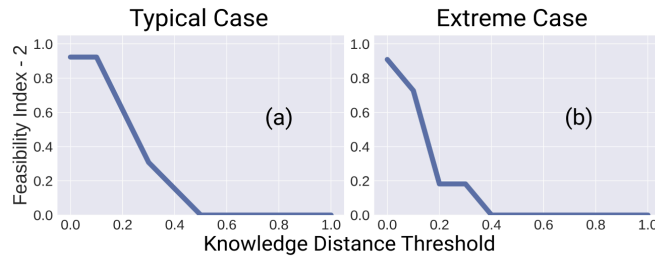


Figure 3.2.9. Policy 3 for a Typical (a) and Extreme (b) Case

Extreme Case. As shown in Figure 3.2.9 (b), for the extreme case (Session 1, C5), while the downward trend was similar, I observed a more rapid decrease as compared to the typical case. Specifically, the FI_2 dropped to only 20% when the KD threshold was as low as 0.2, compared to 60% of FI_2 at the same KD threshold in the typical case. This comparison indicated that some class sessions were more heavily influenced by the parameter of the required knowledge distance threshold, and the effect may differ from session to session.

Policy 4. From the previous analyses, I noted that Policy 4 performed reliably and similarly across classes, making it harder to select an extreme case or a worst-case scenario. I selected a session where Policy 4 performed surprisingly well (Session 2, C3). In Figure 3.2.10, I visualized Policy 4 simulation on two contrasting cases, plotting FI_2 for every step of the knowledge distance ceiling γ for that session.

Typical Case. The tradeoff between knowledge homogeneity and policy feasibility was less prominent than under Policy 3. This means that instructors can afford to choose a stricter (i.e., lower) ceiling so students have a very small knowledge distance, and still achieve high feasibility (FI_2). For example, in Figure 3.2.9 (a), I saw that even if the instructor chose a very strict threshold of $\gamma = 0.1$, nearly 95% of students were able to find a potential partner.

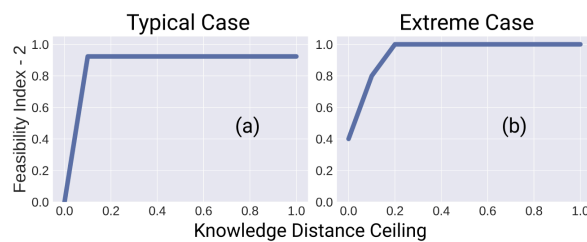


Figure 3.2.10. Policy 4 for a Typical (a) and Extreme (b) Case

Extreme Case. In Figure 3.2.10 (b), even when the KD ceiling was set to 0 (which means students must have the same level of mastery to be paired up), 40% of the students can still be paired, unlike typical cases where usually no two students have the exact same IK. Another noteworthy distinction is that, while the typical case did not reach $FI_2 = 1$ with any

ceiling of KD, the extreme case successfully paired all students ($FI_2 = 1$) with a relatively low ceiling of 0.2.

3.2.7 Discussion

In line with previous LGF research (Maqtary et al., 2019), this work introduces four dynamic LGF policies contextualized in ITS and grounded in user research with K-12 teachers (Echeverria et al., 2020; J. K. Olsen et al., 2020). In this section, I discuss the main findings for research questions, grounded design recommendations for pairing orchestration tools, and future research directions for dynamic LGF.

Main Findings for Research Questions. The data come from middle school classrooms in which students used an intelligent tutoring system to practice equation solving. Regarding the feasibility of the pairing policies (RQ1) and how the feasibility may depend on parameters of the pairing policies (RQ2), I found that, *averaged across time and sessions*, it is generally feasible (93.6%) to team up struggling students with non-struggling students, the minute they struggle (Policy 1). This result remains true even when a high percentage of students is deemed ineligible for being teamed up with struggling students (Policy 2). Specifically, the average feasibility remains above 80% of struggling students across all sessions unless the pairing restriction rate is above 40%. However, as I see in the case study, there can be sessions and moments with high struggle ratios (hence, low feasibility) when using Policy 1. Relatedly, sessions with very high struggle seem more susceptible to the influence of the PR rate in Policy 2 than a typical session.

When pairing students based on whether their knowledge levels are different (Policy 3) or similar (Policy 4), the policy feasibility is highly dependent on the required KD. For Policy 3, there is a tradeoff between the desired heterogeneity (i.e., the knowledge distance threshold) and the policy's feasibility. This means instructors cannot set a high threshold for the KD if they want to pair most students. In Policy 4, the corresponding tradeoff (between homogeneity in knowledge and policy feasibility) is less prominent. Instructors may choose a stricter ceiling for students' similarity in knowledge levels and still achieve high policy feasibility. In the case study, I found that the policy feasibility in different sessions can be differently influenced by the required KD threshold or ceiling, depending on how closely clustered together the students' IK is. For a given, fixed KD threshold (ceiling), a class of students with closely clustered IK may result in higher feasibility for Policy 4 and lower feasibility for Policy 3. Presumably, the feasibility of these policies also depends on class size. For example, from my analysis, I hypothesize that for larger classes, the feasibility of pairing policies may change less drastically when the policy parameters change or as the class progresses.

Regarding policy feasibility by class (RQ3), results show no significant difference among classes for Policy 1 (Struggle with non-struggle) or Policy 4 (Knowledge similarity pairing).

However, I observed significant differences among classes for Policy 2 (Pairing with restriction) and Policy 3 (Knowledge difference pairing). Although different classes have significantly different initial knowledge and struggle status, these differences in IK and struggle status are not always correlated with the feasibility of policies for that class. For example, classes that have different IK may not always have different feasibility for Policy 3 or 4.

Recommendations for Tool Design. The current study aims to inform the design of an orchestration tool that can help pair students dynamically. I aim to lessen teachers' orchestration load when managing *fluid social transitions*. Such a tool plays a key role in my vision for the smart classroom of the future, in which students alternate fluidly between individual and collaborative learning. Here, I highlight three design implications grounded in findings from the current work. These design implications may inform tools that aim to help teachers manage fluid social transitions and ensure the feasibility of dynamic LGF policies. It may also offer inspiration, more broadly, for orchestration tools that aim to team up students in CSCL.

Firstly, technology could be used to automatically adjust the parameters used in LGF policies. My study suggests that the four pairing policies provide a promising foundation for an orchestration tool, but greater flexibility is needed to deal with a wider range of circumstances than each individual policy provides. While some policies (e.g., Policies 1 and 2) explored in this study have a good chance of working well during many class sessions, any given instantiation of a policy (with fixed parameter settings) does not fully deal with class variability and extreme cases. One way to compensate might be to have the tool automatically loosen policy parameters as needed. For example, the tool may gradually loosen the KD threshold or ceiling for policies 3 and 4, when it senses the pairing feasibility to be low.

Secondly, technology could use multiple LGF criteria in a cascading fashion to achieve high feasibility. Specifically, the tool may start out using the ideal pairing policies, and then iteratively try “looser” criteria if the previous one fails to pair up all students. For example, the tool may first attempt to team up students based on struggles on specific KCs - a criterion that is more specific (and restrictive) than Policy 1, but one that could potentially be more effective for helping struggling students. If that fails, then it might pair up students based on their general struggle (Policy 1). If that fails again, then the tool could try to pair students based on knowledge distance. The tool could also customize its pairing criteria, such as using students' characteristics that make the most sense in the given classroom context.

Lastly, technology could be used to recommend LGF policies or policy parameters of high feasibility to teachers. Instead of relying solely on the teachers to make pairing decisions, the tool may adopt *SimPairing* to automatically calculate and maximize policies' feasibility

based on classroom contexts and recommend them to teachers. For example, if the tool determined, using historical transaction data, that students in a given class have consistently low struggling ratios and fewer wheel-spinning students than non-wheel-spinning ones, it may advise that teachers adopt Policy 1 as it has high feasibility. In addition, my findings open up the potential for the tool to help teachers make informed decisions about parameter configuration by notifying them of expected feasibility. For example, if a teacher severely restricts the acceptable pairings, the tool could alert teachers to the low feasibility of the pairing policy and ask if the teacher might want to loosen the restrictions. Running such simulations and providing notification can inform the teachers about the outcomes of the policy's feasibility in their classroom, prior to implementing it. This may prevent teachers from choosing policy configurations that are misaligned with their goals.

Future Research Directions of Dynamic LGF. Building on my investigation, I outline four potential directions for how future research could further explore dynamic LGF policies.

Firstly, in addition to inter-session pairing based on knowledge distance (Policies 3 and 4), future work could explore *intra-session grouping based on knowledge level*, which allows forming pairings during the learning process. Researchers should explore how it would differ from inter-session grouping and which approach better supports teachers' needs. *Secondly*, the current policies identify the students as wheel-spinning if they struggle on *any* of the KCs. Future work could explore whether teachers prefer to pair students based on their KC-specific struggle status. For example, to help a student struggling on the KC *combine constant terms* in equation solving, teachers may prefer to find a partner who has already mastered the same KC, or at minimum is not struggling on the same KC; they may (or may not) find it acceptable, if the partner is struggling on another KC, e.g., *divide by variable coefficient*. Relatedly, teaming up students struggling with different knowledge components may have benefits. Such a pair of students may have complementary knowledge and strengths, and may help each other get unstuck and stop wheel-spinning. Such a pairing criterion opens up good opportunities for tutor-tutee role-switching and mutual peer tutoring. *Thirdly*, analogous to pairing based on KC-specific struggle status, instead of using the mean of students' mastery on different KCs to represent their knowledge, future work could explore to what extent *pairing students based on KC-specific knowledge distance* can be more effective, feasible, or preferable for teachers. KC-specific knowledge pairing might be useful for Policy 3, if teachers want two students who have very different skill levels on one specific KC, so that the one with higher mastery on that KC can tutor the one with lower mastery. *Lastly*, in addition to knowledge level and struggle status, which this work investigated for dynamic grouping, future work can investigate *other student characteristics* (e.g., history of collaborative episodes, preferences for working individually or collaboratively) or *other sources for knowledge level* (e.g., exams or quiz scores) for dynamic pairing. It may also be

especially promising to pair based on dynamic student behaviors detected in real-time by ITS from interaction data, to allow fluid social transitions and dynamic pairing.

3.2.8 Study Contribution and Connection to Thesis Theme

I studied the feasibility of pairing policies in the context of ITS, to inform the design of a tool for orchestrating fluid transitions between individual and collaborative learning. My findings showed that, on average, dynamically pairing students based on their in-the-moment wheel-spinning status resulted in good pairing feasibility for struggling students, even with moderate restrictions on the allowed pairings. I found there was trade-off between the *required knowledge distance* and *the policy feasibility*, meaning the stricter the required knowledge distance is, the less feasible a pairing policy can be. Such a trade-off was more prominent in heterogeneous grouping (different knowledge levels) than in homogeneous grouping (similar knowledge levels). However, any given instantiation of a policy (with fixed parameter settings) did not fully deal with class variability and extreme cases, as policies have different feasibility for different classes and sessions. This suggested optimization for policy feasibility (e.g., through gradually loosening parameters) or classroom customization needed to be taken into consideration. Methodologically, this research extends previous work (e.g., Replay Enactments) that used authentic data and algorithms as design materials to augment designers' intuitions for designing future tools (Holstein, Harpstead, et al., 2020). This work had several novel elements. First, using the *SimPairing* approach, I explored the *feasibility* of LGF policies derived from user research with math teachers. I addressed a gap in the literature for dynamic intra-session LGF (Amara et al., 2016) and envisioned how instructors and/or an orchestration tool would customize pairing policies and parameters to specific classroom contexts, which prior work argued to be especially helpful in the LGF process (Amara et al., 2016; Maqtary et al., 2019).

In sum, theoretically, this work bridged the literature gap on its investigation of the *feasibility* of user-centered dynamic pairing policies. Practically, I contributed grounded design directions for pairing orchestration tools and *SimPairing* as an approach to evaluate dynamic LGF policies, which may generalize to other online educational software that has transaction data. This work connects to making the AI-based technology more customized to classrooms because: 1) it studies how different classroom contexts may have different feasibility for pairing policies, 2) It shows that different parameters may be tuned to different classroom contexts so that it ensures a desirable paired-up proportion for different pairing policies. Unlike Chapter 3.1, which is about customized instructional materials, the broad theme of this work is mainly connected to customized technology (specifically, dynamic pairing algorithms) for classroom contexts, based on empirical exploration of data simulation results.

Additionally, this piece of work on dynamic pairing policies was done to build a co-orchestration tool for dynamic pairing. This work enables us to make the design decision in the co-orchestration tool Pair-Up (to be explained later in Chapter 4), especially on how teachers choose pairing policies. Connecting this work and Chapter 4.1, where teachers expressed preferences for making final decisions, the practical results and design realization in the Pair-Up tool is that teachers could select which pairing algorithm they would like to use for their specific classroom contexts. While the Pair-Up currently does not have parameter tuning functionality (for simplicity of the tool), it could be a viable future option, especially for technology-savvy teachers or teachers who are already familiar with the basic version of the pairing tool and would like to customize further.

II. DIFFERENTIATION

In this part, the studies I grouped relate to the “differentiation” thesis theme. There are many ways to make learning environments allow for differentiated and personalized learning. Prior work has demonstrated that combining individual and collaborative learning may be more helpful than either mode alone (J. K. Olsen et al., 2019). My thesis specifically focuses on dynamically combining different learning modes for students, where students switch back and forth between individual and collaborative learning. This constitutes a form of differentiated learning because students can engage in different learning activities in a self-paced way.

My work focuses on answering several main research questions:

1. How can I build a technology ecosystem that leverages AI and teachers’ strengths, respects teachers’ preferences, and supports a dynamic combination of individual and collaborative learning?
2. Is it feasible for teachers and students to achieve these combinations with technology support?
3. How do dynamic combinations of individual and collaborative learning compare to traditional classroom pairing?

I conducted a storyboard survey to address these questions and understand teachers' preferences and boundaries in collaborating with AI for dynamic transitions. Working with the research team, I constructed a co-orchestration technology ecosystem, including intelligent tutoring systems for individual and collaborative learning and a teacher-facing tool named Pair-Up (Studies 3 and 4 in Table 1.1). Pair-Up supports human-AI co-orchestration of students’ transitions between learning modes, displaying real-time learning analytics and offering pairing suggestions while preserving teacher agency.

I conducted a classroom-based prototyping study of dynamic transitions (Study 5, Table 1.1) and a controlled study comparing standard versus dynamic transitions (Study 6, Table 1.1). These studies aimed to create and evaluate dynamically differentiated learning environments supported by intelligent tutoring systems. The design respected teachers' preferences for boundaries and agency in AI collaboration.

CHAPTER 4 USER RESEARCH ON TEACHER PREFERENCE AND BOUNDARY IN HUMAN-AI CONTROL (IN CONTEXT OF DYNAMICALLY DIFFERENTIATED LEARNING BETWEEN INDIVIDUAL AND COLLABORATIVE LEARNING)

Acknowledgement. This study was done collaboratively with the post-doc on the project LuEttaMae Lawrence, a previous post-doc (who contributed to the visualization of storyboard and mentoring of the study procedure), Vanessa Echeverria (who has left the post-doc position at that time but remained a collaborator), Bill Guo (undergraduate research assistant) and under the supervision of PIs (Vincent Aleven and Nikol Rummel).

A collaborative ITS, APTA, extends Lynnette’s functionality to support reciprocal peer tutoring. Both tutoring systems’ initial build and design were not part of my thesis work. Lynnette was built and designed primarily by the research team led by Vincent Aleven. APTA was initially the PhD work of Prof. Erin Walker (supervised by Kenneth Koedinger and Nikol Rummel). While I have, during my PhD, contributed to the re-design of APTA, the re-design was for the purpose of improving user experience and clarity in workflow (which I do not describe in the thesis). In this thesis, when mentioning studies on these two tutoring systems, I either used them as platforms for conducting studies or contexts for user research studies, and do not claim intellectual credit for designing or building the systems.

Chapter 4 is based on study 3 in Table 1.1 and primarily on this paper⁴.

4.1 Study Summary

Orchestration tools may support K-12 teachers in facilitating student learning, especially when designed to address classroom stakeholders’ needs. My previous work revealed a need for human-AI shared control when dynamically pairing students for collaborative learning in the classroom, but offered limited guidance on the role each agent should take. In this study, with the help of colleagues, I designed storyboards for scenarios where teachers, students, and AI co-orchestrate dynamic pairing when using AI-based adaptive math software for individual and collaborative learning. We surveyed 54 math teachers on their co-orchestration preferences. I found that teachers would like to share control with the AI to lessen their orchestration load. As well, they would like to have the AI propose student pairs with explanations and identify risky proposed pairings. However, teachers are hesitant to let the AI

⁴ Kexin Bella Yang, LuEttaMae Lawrence, Vanessa Echeverria, Boyuan Guo, Nikol Rummel, Vincent Aleven. “Surveying Teachers’ Preferences and Boundaries regarding Human-AI Control in Dynamic Pairing of Students for Collaborative Learning”. (*EC-TEL’ 21*, [PDF](#))

auto-pair students even if they are busy, and are less inclined to let AI override teacher-proposed pairing. My study contributes to teachers' needs, preferences, and boundaries for how they want to share the task and control of student pairing with the AI and students, and design implications in human-AI co-orchestration tools.

4.2 Background and Motivation

While teachers need to facilitate students' collaborative activities and monitor their progress, these activities can be cognitively demanding for teachers (Hämäläinen & Vähäsantanen, 2011; Olsen et al., 2020; Santos, 2012; van Leeuwen & Rummel, 2019). Orchestration broadly refers to the planning and real-time management of learning activities in the classroom (Holstein and Olsen, 2021). Various orchestration tools and prototypes have been designed to help lessen teachers' load of class management or allow them to focus on teaching and assisting students (Holstein et al., 2017; J. K. Olsen et al., 2020; Santos, 2012; Sharples, 2013; Tissenbaum & Slotta, 2019). Recently, much work focuses on "co-orchestration," referring to technology that is explicitly designed so that the orchestration responsibilities are shared across multiple agents (e.g., teachers, students, and AI systems) (Holstein & Olsen, 2021; Santos, 2012; Sharples, 2013).

Much research has focused on designing tools to support teachers orchestrating individual (e.g., Alavi et al., 2009) or collaborative learning scenarios (Martinez-Maldonado et al., 2015; van Leeuwen & Rummel, 2019). These tools have typically been designed assuming that students progress through activities in a pre-planned, relatively synchronized manner. By contrast, little work has focused on supporting fluid social transitions (with recent exceptions (Echeverria et al., 2020; J. K. Olsen et al., 2020; Tissenbaum & Slotta, 2019)). Fluid social transitions refer to transitions between classroom activities that "occur asynchronously between students - not all at the same time for everyone in the class" (J. K. Olsen et al., 2020). Unlike the static, planned transitions, fluid social transitions may be needed in technology-enhanced classrooms to support students to flexibly move between activities at a pace that suits their specific circumstances and knowledge level. It is a hypothesis that technologies that support fluid transitions can better support students' learning (J. K. Olsen et al., 2020). To be able to test this hypothesis, the high-level goal for the current study is to design a co-orchestration tool to help teachers easily and effectively manage fluid social transitions in class. Specifically, I focused on dynamic pairing, which means teaming up students opportunistically based on unfolding learning situations (Echeverria et al., 2020; J. K. Olsen et al., 2020).

To design effective tools that can help K-12 teachers co-orchestrate their students' learning activities in classrooms, it is critical that the tools "support the needs and respect the boundaries of both teachers and students" (Holstein et al., 2019). Researchers have begun

exploring questions such as: How should orchestration responsibilities be divided among different classroom stakeholders? Who should be accountable for particular instructional decisions? How should such hybrid control adapt to context and learning scenarios (Echeverria et al., 2020; Holstein et al., 2019; Kenneth Holstein And Jennifer K Olsen, 2021)? Our previous work reveals that classroom stakeholders have nuanced preferences regarding co-orchestrating classroom activities. From design research with K-12 teachers, researchers found “a delicate balance between automation and respecting teachers’ autonomy”, and that “over-automation risks threatening teachers’ authority in class and flexibility to set their own goals, yet under-automation may burden teachers with tasks they’d rather not perform” (Holstein et al., 2019). Similarly, prior work has found from design research with primary school teachers that instructors prefer to “maintain an elevated position above AI systems and need some degree of accountability and control” (J. K. Olsen et al., 2020). Echeverria et al. found from a Wizard-of-Oz technology probe in K-12 classrooms a need for hybrid control between students, teachers, and AI systems over dynamic transitions from individual to collaborative learning. They also found a need for such hybrid control to be adaptable to classroom contexts, such as class size and students’ prior knowledge.

This study investigates the following questions: In co-orchestrating dynamic student pairing, how do teachers want to share control with AI systems and students, regarding proposing, evaluating, and deciding pairings? (RQ1) What criteria do teachers prefer when dynamically pairing students? (RQ2) In the current study, I designed storyboards of possible scenarios of human-AI co-orchestration of dynamic pairing. I surveyed 54 math teachers’ preferences on them, in the context of using adaptive AI-tutoring software for individual and collaborative K-12 math learning.

4.3 Methods and Study Design

The overall goal of the research project, which the current study is part of, is to design a co-orchestration tool that helps teachers manage fluid transitions back and forth between individual and collaborative learning in the classroom. This study concerns finding teachers’ needs and preferences on one key aspect in managing fluid transitions: dynamically teaming up students (i.e., dynamic pairing). Dynamic pairing may not happen at the same time for every student, and managing the process in real-time may be overwhelming for teachers.

Many design-based research activities, such as user interviews, can require significant time input from researchers, often resulting in a relatively smaller participant pool. Surveys, while being easily scalable, are often hard to adequately convey the nuanced designs to the participants. To concisely communicate potential designs to a larger user population, I incorporated storyboards into a survey. I created storyboards that illustrate different co-orchestration tool features in the context of different dynamic scenarios. I surveyed

teachers’ opinions on these scenarios to make sure the tool can be aligned with teachers’ needs and preferences. These scenarios take context in two intelligent tutoring systems (ITSs) that support learning of middle school math algebra, specifically, equation solving. The individual ITS, Lynnette, has been proven to improve students’ equation-solving skills in several classroom studies (e.g., (Holstein, McLaren, et al., 2018)). The collaborative ITS, APTA, extends Lynnette’s functionality to support reciprocal peer tutoring. The initial build and design of both tutoring systems were not part of my thesis work. Lynnette was built and designed primarily by the research team led by Vincent Aleven. APTA was initially the PhD work of Prof. Erin Walker (supervised by Kenneth Koedinger and Nikol Rummel). While I have, during my PhD, contributed to the re-design of APTA, the re-design was for the purpose of improving user experience and clarity in workflow (which I would not describe in the thesis). In this thesis, when mentioning studies done on these two tutoring systems, I either used them as platforms for conducting studies or contexts for user research studies, and do not claim intellectual credit for either designing or building the systems.

Using storyboards is a standard method in human-centered design. Storyboards show how users interact with different versions of a proposed system in specific contexts (Davidoff et al., 2007). According to Davidoff et al., in the need validation process, storyboards may “help designers prioritize user needs, more clearly map spaces for innovation, and use that focus to narrow the design space for potential applications.” (Davidoff et al., 2007)

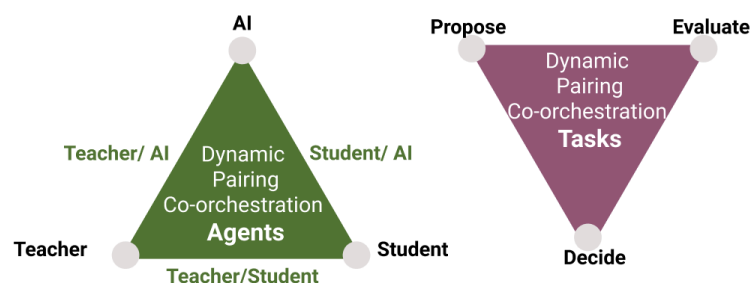


Fig 4.1.1. Agents (left) and Tasks (right) in Human-AI Orchestration of Dynamic Student Pairing

Co-orchestration Scenarios and Survey Design. To inform the design space, the previous post-doc (LuEttaMae Lawrence) and I reviewed literature at the intersection of orchestration tools, dynamic pairing, fluid social transitions, and computer-supported collaborative learning (CSCL) (Echeverria et al., 2020; Holstein, Aleven, et al., 2020; J. K. Olsen et al., 2020; Sankaranarayanan et al., 2020; van Leeuwen et al., 2018). As shown in Fig. 4.1.1, for co-orchestrating dynamic pairing, there are three main agents and three main tasks. The three main agents are the teacher, AI, and student (Fig. 4.1.1, left), and the three main functions are how the student pairings are proposed, evaluated, and decided (Fig. 4.1.1, right). The design space for dynamic pairing co-orchestration is about which agent(s) have control over each

task. Two authors collaboratively brainstormed 22 design concepts for potential co-orchestration scenarios. I used the past research on teachers' needs in pairing co-orchestration (Echeverria et al., 2020; J. K. Olsen et al., 2020) to prioritize scenarios teachers may find more useful. I also included scenarios that push social boundaries or may be controversial among teachers, as it may help to uncover where these boundaries lie (Davidoff et al., 2007). I went through four rounds of clustering and refining the 22 design concepts, and finalized five co-orchestration scenarios, described in Table 4.1 (first column).

Based on these five scenarios, we (the post-doc LuEttaMae Lawrence and I) designed a survey that had three sections: 1) teachers' demographics and teaching experience, 2) five co-orchestration scenarios and related questions, and 3) teachers' general preferences on pairing co-orchestration. The five scenarios were presented in a random sequence for every participant, to reduce bias and carry-over effects. Each scenario has 1) a three-panel storyboard with a short title with a simple visual and description (e.g., Fig. 4.1.2), 2) a seven-point Likert scale question asking how likely it is that the teacher would use the technology shown in the storyboard in their classroom, 3) a follow up open-ended question asking why or why not, 4) an open-ended question asking what improvements or changes teachers would want to make to the scenario, 5) 1-2 focused seven-point Likert question specific to the main design element in the scenario. In addition, some scenarios had 6) 1-2 multiple-choice probing questions. The complete survey questions can be found in supplementary materials.

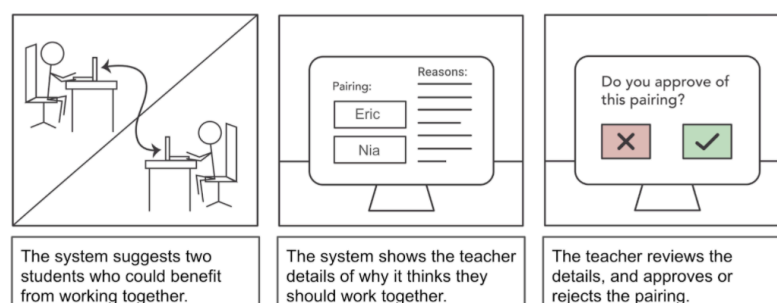


Fig. 4.1.2. Storyboard Example (Co-orchestration Scenario 3)

Procedure. The survey was hosted online using Qualtrics and was available for seven days. Teachers were introduced to the context and asked to consent to the research. Before large-scale deployment, researchers conducted eight pilot sessions, including two think-aloud studies, with experienced math teachers and researchers. Based on pilot results, the survey and storyboards were iteratively refined for clarity.

Participants. I recruited 54 math teachers (36 females, 18 males) from social media groups. They completed a 20-30 minute survey and received a \$10 USD Amazon Gift Card. Most teachers (75%) taught in middle school and had over 4 years of experience (88%). The majority were white (80%) and taught in the USA (96%). At the time, 52% taught fully online, 41% in hybrid mode, and 4% fully in-person. Most (85%) used collaborative learning activities at least 50% of the time. Other demographics are in the supplementary materials.

Table 4.1 Co-orchestration Scenario Descriptions and Teachers' Stated Likelihood of Use

Scenario Titles and Descriptions	Co-orchestration Roles	Stated Likelihood of Use	
		M (SD) ⁵	Median
<i>S1. Teacher Informs Automated Pairing:</i> Teacher entered information for the AI to use when making student pairings.	<i>Proposes:</i> Teacher and System <i>Evaluate:</i> N/A <i>Decides:</i> AI	4.96 (1.58)	5
<i>S2. AI Auto-Pairs When Teachers are Busy:</i> AI paired students up on its own because the teacher was busy and in "Do not disturb" mode.	<i>Proposes:</i> AI <i>Evaluate:</i> N/A <i>Decides:</i> AI	4.70 (1.59)	5
<i>S3. AI Explains Pairing Suggestions:</i> AI paired students up, explained to the teacher why it paired them up and asked the teacher to approve or reject the pairing.	<i>Proposes:</i> AI <i>Evaluate:</i> Teacher <i>Decides:</i> Teacher	5.89 (1.13)	6
<i>S4. AI Identifies Risky Pairs:</i> Students proposed a partner, the AI reviewed the pairing and notified teachers of potentially risky pairings for teachers to approve or reject.	<i>Proposes:</i> Student <i>Evaluate:</i> AI + Teacher <i>Decides:</i> Teacher	5.33 (1.41)	6
<i>S5. AI Reviews and Decides:</i> Teachers paired students up, the AI evaluated the pairings, and changed risky ones without notifying teachers.	<i>Proposes:</i> Teacher <i>Evaluate:</i> AI alone <i>Decides:</i> AI	4.00 (1.76)	4

Data Analysis Approach. I manually reviewed and filtered out duplicate or invalid survey responses. For quantitative analysis, I computed the mean and SD for seven-point Likert scale questions and analyzed responses to multiple-choice questions. For qualitative analysis, I examined 540 open-ended responses to questions about technology use and desired improvements. Using iterative affinity diagramming, two researchers grouped responses into 117 first-level and 63 second-level themes. To synthesize findings, I laid out analysis results within each scenario to understand teachers' preferences and sentiments on design elements in the co-orchestration tool. I organized quantitative results and qualitative themes by research questions to formulate insights regarding teachers' co-orchestration preferences.

⁵ Likert Scale Labels: 1 - Definitely No, 2 - Very Unlikely, 3 - Probably not, 4 - Neither likely nor unlikely, 5 - Probably yes, 6 - Very likely, 7 - Definitely yes

4.4 Results

Scenario 1 Results: In Scenario 1 (S1), teachers entered information about each student for AI to use in making pairings. S1 was the third-highest in terms of teachers' likelihood of use ($M = 4.96$, $SD = 1.58$). When asked about entering student information, 37% of teachers were "very likely" or "definitely yes" ($M = 4.65$, $SD = 1.85$). Most teachers (70%) preferred personalized pairing criteria. The most important factors for pairing were students' knowledge mastery (98%), error rate (76%), pairing history (76%), personality (66%), number of collaborative sessions (64%), friendships/relationships (52%), and gender (34%). Qualitative analysis showed teachers liked the potential efficiency and reduced orchestration load ($N = 7$) and appreciated the flexibility of front-loading preparation work ($N = 6$). However, concerns about the time-consuming nature of entering information ($N = 12$), especially for large classes, and privacy ($N = 2$) were noted. Teachers suggested automating data collection or using self-assessments ($N = 7$).

Scenario 2 Results: In Scenario 2 (S2), the AI auto-pairs students when teachers activate a "Do not disturb" mode. S2 had the second-lowest likelihood of use ($M = 4.7$, $SD = 1.59$). Only 37% of teachers agreed that the AI should pair students without their approval when busy ($M = 4.35$, $SD = 1.78$). Most preferred AI pairing when helping other students (57%), during off hours (22%), or never (17%). Teachers felt auto-pairing could reduce distractions and improve efficiency ($N = 19$). However, many ($N = 32$) had reservations about AI control and preferred to review and potentially override pairings ($N = 26$). Concerns included undermining teacher authority and the complexity of the feature. Teachers suggested unobtrusive notifications or quick approvals.

Scenario 3 Results: In Scenario 3 (S3), the system proposes pairings with explanations for teacher approval or rejection. S3 was the most favorable scenario ($M = 5.89$, $SD = 1.13$). A significant majority (77%) wanted the system to explain its reasoning ($M = 6.07$, $SD = 0.93$). Key information desired included students' math skill mastery (94%), the problems each student is working on (92%), and recent errors (88%). Teachers appreciated the idea of system-suggested pairings with explanations ($N = 38$), increasing work efficiency ($N = 10$), and maintaining control over final decisions ($N = 11$). Some suggested further improvements, like finding alternative partners and ensuring accessibility.

Scenario 4 Results: In Scenario 4 (S4), students suggest partners, the AI reviews and flags risky pairings for teacher approval. S4 had the second-highest likelihood of use ($M = 5.33$, $SD = 1.41$). Most teachers (70%) agreed that the AI should notify them about risky pairings ($M = 5.72$, $SD = 1.19$). Preferences varied on how to notify students when pairings are rejected. Teachers liked involving students in pairing decisions ($N = 16$) and the AI's role in detecting risky pairings ($N = 13$). Concerns included the potential negative impact on

teacher-student relationships when rejecting pairs ($N = 13$). Teachers wanted to see the pairing history and results ($N = 9$).

Scenario 5 Results: In Scenario 5 (S5), teachers propose pairings, and the AI changes risky pairings without notification. S5 had the lowest likelihood of use and highest controversy ($M = 4.00$, $SD = 1.76$). While 47% agreed the AI should review pairings ($M = 4.87$, $SD = 1.64$), only 15% agreed it should override them without notification ($M = 3.39$, $SD = 1.86$). Teachers were divided; some liked the additional oversight ($N = 7$) and efficiency ($N = 10$), but most wanted control over final decisions ($N = 53$). They preferred notifications and the ability to override AI decisions ($N = 21$).

Teachers' Other Preferences on Hybrid Control in Pairing. I report how teachers want to share the control of *proposing* and *deciding* with AI and students (RQ1), as stated in the *general preferences* section of the survey. Regarding *who should propose pairings*, most teachers (90%) think *teachers* should be involved. Interestingly, more teachers thought the *system* (76%) should be involved than *students* should be involved (53%). In addition, 80% of teachers thought that suggesting or proposing student pairing should be *shared*. Among them, the two most popular co-orchestration choices were sharing control *between teacher, student, and system* (41%), or *between teacher and system* (30%). Regarding *who should make the final decisions* about pairings, 95% of the teachers thought *teachers* should be the ones to do so. Only two teachers (4%) thought students should make the final decision about student pairing, and only one thought the system should.

4.5 Discussion

Insights Related to the Research Questions. I discuss the roles teachers think each agent (i.e., teacher, AI system, students) should or should not take when co-orchestrating dynamic pairing (RQ1), and their preferences regarding dynamic pairing criteria (RQ2).

Teachers' Role. Across all scenarios, teachers prefer to prepare for student pairing before class, make final pairing decisions, and customize the orchestration tool. Firstly, teachers want to contribute knowledge to help pair up students and to front-load such preparation work prior to class. They want to set pairing restrictions (i.e., identify students who should not be teamed up) and pre-approve pairings. However, many teachers, especially those who teach large classes, do not want to enter information into the system to help the AI pair students, as it may raise privacy concerns or may be too time-consuming. Given that the survey question did not specify either the type or the amount of student information teachers would need to enter, it seems possible that many teachers might be willing to enter a small amount of information, in line with Olsen's finding from co-design studies with teachers (J. Olsen, 2017). Thus, a simple and time-efficient design that allows teachers to set restricted and pre-approved pairings may be needed. Secondly, teachers want to be able to review and

modify student pairings proposed by other agents; they strongly want to have the final say. Thirdly, teachers want to be able to customize and configure the co-orchestration tool (e.g., pairing criteria, frequency of changing pairs) to fit their classroom context.

AI Systems' Role. Teachers like the AI to propose personalized pairings, explain the reasoning behind proposed pairings, help evaluate proposed pairings, and lessen teachers' orchestration load. They do not want the AI to make final pairing decisions. Teachers like the system to propose pairings that are personalized to each student's characteristics. The top three factors that teachers think the system should consider in dynamic pairing are students' knowledge mastery, overall error rate, and past pairing partners. Teachers consider students' friendships, relationships, and gender to be less important in pairing. Most teachers want to use different criteria for different students, suggesting that students' characteristics may be weighed in the pairing decision in a manner that varies by situation.

Additionally, teachers want to see reasoning and explanations for why the AI proposes to pair two students or considers a pair to be risky. Teachers want the AI system to act as an extra pair of eyes to review and evaluate proposed pairings, no matter if they are student-proposed or teacher-proposed. Furthermore, teachers want the AI system to lessen their orchestration load, increase work flexibility (e.g., by allowing them to front-load pairing preparation tasks), and minimize distractions while working with students. Interestingly, while teachers want to reduce students' time waiting for their help, they do not all agree that the AI should auto-pair students even when they are busy. Though teachers like being assisted by the AI, most teachers rejected the idea for the AI to pair students without teachers' review or approval.

Students' Role. Teachers see value in allowing students to have a say in the pairing process and provide feedback on pairing, as "students are more likely to be productive if they are given the opportunity to have a sense of ownership in their partnership" (T23). One teacher mentioned it may be helpful to notify teachers if "one partner was unwilling to work [with another student]" (T41). Compared to teachers' and AI's roles, teachers made fewer comments about the role that students should have, which may be explained by the fact that I have more scenarios describing teacher-AI co-orchestration than student-AI or teacher-student-AI.

4.6 Design Challenges and Directions

Design Challenges. My findings uncover several design challenges. First, teachers' wish to reduce students' waiting time and avoid being the bottleneck blocking students' progress suggests that some degree of system autonomy may be needed. The design challenge is to do so without letting teachers feel their authority in class is threatened. Secondly, there may be a tension between teachers' desired awareness and control, and their desires to avoid interference and distraction when working with students. Relatedly, since teachers often need

to make fast decisions when teaching and managing classroom activities, there may be limited time for them to “consume and digest” the explanations systems give. It may be necessary to investigate what information will be most helpful to teachers in these explanations and create a design that provides interpretable explanations at a glance. These explanations should take teachers minimal time to read and interpret, and still give teachers enough evidence to support their decisions (e.g., approving or rejecting pairings). Finally, it may be hard for a co-orchestration tool to allow students to have a say in pairing decisions and ensure teachers have the final say while avoiding harming teacher-student relationships if students’ proposals are rejected.

Design Directions. In connection with the three tasks in dynamic pairing - propose, evaluate, and decide pairing (Fig. 4.1.1, right), this study presents three design directions, including having the AI system be a teacher’s helper that explains its reasoning (Propose), be an extra pair of eyes (Evaluate), and notify instead of deciding (Decide).

Firstly, it may be fruitful for the AI system to monitor the class and help teachers prioritize their attention to those who need it the most. For example, if the AI can keep track of students’ working progress and pairing needs, and display the classes’ status to the teacher in “a queue with time of any students not already being paired” (T8), it may help teachers to prioritize so that students get help sooner than when no AI-monitor is involved. A promising direction is for the AI helper to explain its reasoning when proposing or suggesting certain educational decisions.

Secondly, while teachers want the final say in educational decisions (e.g., student pairing), they want the AI to provide an extra pair of eyes, capture things that may escape teachers’ consideration, and augment teachers’ memory as a form of “distributed cognition.” For example, given that an AI can track student pairing history, it may suggest to teachers whether or not two students should be paired again. The AI can base its suggestions on the total number of times students have been paired and perhaps their past collaboration quality, information that teachers may not readily have.

Lastly, this study suggests that it may be fruitful to use the AI systems to notify teachers of worrying classroom activities that are worth their attention (e.g., risky pairings or unproductive students), and ask them to make final decisions. This may prove much more preferable for teachers than if the AI makes decisions, as it ensures teacher awareness and control. More iterative work is needed to design the orchestration tool, attending to both teachers’ preferences and the practical feasibility of different pairing policies (Yang, Echeverria, et al., 2021b) for pairing students in a given classroom.

4.7 Study Contribution and How it Connects to the Thesis Theme

Firstly, this study confirmed, with larger sample size, findings from prior design research and classroom studies (Echeverria et al., 2020; Holstein et al., 2019b; J. K. Olsen et al., 2020; Santos, 2012) that teachers want to: 1) have the final say over pairing students, 2) enable students to get help from other sources when teachers are busy, 3) flexibly front-load preparation tasks for collaborative activities *before class*, and 4) minimize interruptions and orchestration load *during class*. Secondly, I uncover new insights into teachers' needs regarding how to share aspects of dynamic pairing with the AI and with students. The insights include 1) teachers want to see brief reasoning given by the system in at least two situations: when the AI system proposes a pairing, and when it detects a risky pairing. 2) Although teachers want to contribute to student pairings, they are concerned that entering student information into the system may be time-consuming. Thirdly, this study explored how teachers' preferences of orchestration control may depend on certain *dynamic contexts*, including when teachers are busy, when risky pairings occur, when the teacher (S5), system (S2,3), or student (S4) proposed pairings.

As researchers start envisioning more sophisticated and personalized interactions in future smart classrooms, fluid social transitions become an interesting issue to study (Yang, Echeverria, et al., 2021b). This work contributes to the literature of orchestration for fluid social transitions and dynamic pairing. Based on results of the user-centered, mixed-method research through surveying 54 math teachers, the current study extends and complements prior research in human-AI co-orchestration by validating teachers' preferences with a larger sample size, revealing new, nuanced, and context-dependent needs and preferences, and proposing design implications. As the community increasingly adopts a co-orchestration lens to leverage human and AI's complementary strengths to achieve synergy, I am hopeful that the teacher needs validated and uncovered in this study can help inform future design and research of tools.

This study particularly connects to the theme of "*Differentiation*", as it serves as an important step in the human-centered design of the later technology ecosystem. Identifying how to divide and distribute the control of pairing action between teachers, students, and AI is an important key step for the design of the technology ecosystem, as this study helps us uncover the important needs and preferences of the main stakeholders (teachers) for the technology ecosystem.

CHAPTER 5 PAIR-UP: SUPPORTING DYNAMIC COMBINATION OF INDIVIDUAL AND COLLABORATIVE LEARNING

Acknowledgements. The co-design of Pair-Up was made possible through the close collaboration of former postdoc Lu Lawrence. The design and development of the tool were carried out by the larger research team, including interns and technical staff, with my contributions in providing design insights, testing, and developing the pairing algorithms. The work is made possible by the efforts of research interns Hongyu Mao and Zijing Lu on data analysis, and of the mentorship and supervision of Professors Nikol Rummel, Vincent Aleven, and Kenneth Holstein, as well as Dr. Vanessa Echeverria.

The design of Pair-Up was undertaken iteratively in phases with increasing fidelity, building on and informed directly by three prior studies: 1) surveying teachers' preferences and boundaries in the human-AI co-orchestration process (Yang, Lawrence, et al., 2021), 2) data simulation of possible pairing algorithms (Yang, Echeverria, et al., 2021b), 3) interview and co-design workshops with teachers (Lawrence et al., 2022). From these studies, I have found that 1) teachers prefer system suggestions and have final control over pairing decisions, 2) it's feasible to pair students based on their knowledge level, (specifically, in-the-moment wheel-spinning status) based on historical log data from intelligent tutoring systems, and 3) specific design features and layout choices teachers desire in a real-time orchestration tool. The specific design process of Pair-Up is documented in the following publications ⁶.

Pair-Up is a web-based tool that displays students in card format (Fig. 5.1, left panel). Pair-up displays students' skills and progress, and allows teachers to choose collaborative assignments for students (Fig. 5.1, right panel). The left panel displays cards representing students working individually (upper part) and collaboratively (lower part). The top-right corner of each card (marked "(a)") shows an icon that represents a student's real-time learning status, indicating whether the student is idling, misusing software, making lots of errors, making many attempts, or doing well. The panel on the right supports the teaming up of students. It displays the skills and progress of a candidate for the Solver role and one for the Tutor role, both selected by the teacher, to help the teacher assess whether these two

⁶ Kexin Bella Yang, Zijing Lu, Vanessa Echeverria, Jonathan Sewall, LuEttamae Lawrence, Nikol Rummel, Vincent Aleven. "Technology Ecosystem for Orchestrating Dynamic Transitions between Individual and Collaborative AI-tutored Problem Solving". (*AIED*' 22 short paper, [PDF](#))

LuEttamae Lawrence, Boyuan Guo, Kexin Bella Yang, Vanessa Echeverria, Jimmy Kang, Vikrant Bathala, Christina Li, William Huang, Vincent Aleven, Nikol Rummel "Co-Designing AI-Based Orchestration Tools to Support Dynamic Transitions: Design Narratives Through Conjecture Mapping" (ICLS' 22, [PDF](#), best design paper)

LuEttamae Lawrence, Vanessa Echeverria, Kexin Bella Yang, Vincent Aleven, Nikol Rummel. How teachers conceptualise shared control with an AI co-orchestration tool: A multiyear teacher-centred design process. (August 2023, British Journal of Educational Technology) <https://doi.org/10.1111/bjet.13372>

students might make a good collaborative team. As well, in this panel, the teacher can choose a collaborative assignment for paired-up students to work on.

To support teachers in monitoring students' learning state, Pair-Up *displays students' learning analytics in real-time*, by means of indicators of their recent learning status (a) in Fig. 5.1), which have been co-designed iteratively with teachers (Lawrence et al., 2022). As shown in Fig. 5.1, displaying learning status on a student card could alert teachers to students' behaviors such as idling, misusing the software, making lots of errors, making many attempts, and doing well. Pair-Up also displays students' progress in problems at hand and their estimated knowledge level on knowledge components (Aleven & Koedinger, 2013) that students are practicing. Teachers can sort students in various ways, such as alphabetically, based on the number of math problems solved, or the status indicators (i.e., idle, misuse, lots of errors, many attempts, doing well). The teacher is in charge of deciding who to team up with and when. The system design is commensurate with design guidelines for human-AI systems (Amershi et al., 2019) and is grounded in prior findings from user-centered research with teachers (Yang, Lawrence, et al., 2021). The tool makes suggestions to teachers of which student might take on the role of Solver or a Tutor, and highlights suggested Solvers in teal and Tutors in purple (Fig. 5.1). The teachers *may choose to follow system pairing suggestions or override them*, and pair students based on their own judgment (Echeverria et al., 2020). Informed by prior results on data simulation of pairing algorithms and surveying of teachers' preferred pairing strategies (Yang, Echeverria, et al., 2021b; Yang, Lawrence, et al., 2021), the tool currently has two built-in pairing suggestion algorithms: *random pairing and pairing students by different knowledge levels*. The latter identifies (as candidates for the Solver role) students who are making slow progress on some of the knowledge components and suggests a partner (in the Tutor role) who can likely help them with the knowledge that they are struggling with, based on students' real-time interaction learning data with systems.



Figure 5.1. Pair-Up: A Teacher-Facing Co-Orchestration Tool for Managing Dynamic Transitions

I conducted a classroom prototyping study to evaluate whether it is feasible for the class to achieve dynamic combinations of individual and collaborative learning, as supported by a technology ecosystem, with the key orchestration tool being Pair-Up. This study was done in collaboration with Vanessa Echeverria (a previous post-doc), Zijing Lu, Hongyu Mao, and under the mentorship and supervision of Co-PIs (Kenneth Holstein, Nikol Rummel, and Vincent Aleven). This chapter 5 describes this study, with more details in the published paper⁷.

5.1 Study Motivation and Context

Classroom teaching is a dynamic and complex job requiring teachers to improvise to deliver the most effective instructional support for students at any given moment. Teachers need to continuously observe each learner's current state to flexibly decide how, when, and who to support (An et al., 2020; Kerry & Kerry, 1997; Tomlinson, 2014). When multiple students need help at the same time, a teacher may not be able to attend to their requests and learning needs. Prior classroom observations found that in such moments, teachers may team students up opportunistically, so that students can get help from their peers (Holstein et al., 2019a, 2019b). I refer to such events as *dynamic transitions* between individual and collaborative learning. Dynamic transitions go beyond the pre-planned team formation strategies common in classrooms, where most or all students start or stop collaboration simultaneously. In dynamic transitions, by contrast, teachers have students transition between learning modes (individual or collaborative) *when the need arises*, to be maximally responsive to the fact that students learn at their own pace. Thus, A teacher must gauge which students need help, identify suitable partners for these students, and monitor their collaboration to decide when it should end (i.e., when the students should revert back to individual learning). It is not easy to manage such dynamic transitions, especially given that teachers may need to simultaneously teach, support other students, and attend to unexpected classroom behaviors (Holstein et al., 2019a, 2019b).

The current study prototypes a technology ecosystem that leverages both human and AI system strengths in classroom management. Various educational tools have been designed and developed to support teaching and learning processes mediated by technology (An et al., 2020; Kharrufa et al., 2017). Some of these tools support *classroom orchestration*, which broadly refers to planning, monitoring, and real-time management of complex classroom activities (P. Dillenbourg & Jermann, 2010; Holstein et al., 2019b). Prior work pointed out that orchestration considers interactions at the level of the broader learning ecosystem, such as a classroom with one or more instructors and groups of learners (Dillenbourg, 2013;

⁷ Kexin Bella Yang, Vanessa Echeverria, Zijing Lu, Hongyu Mao, Kenneth Holstein, Nikol Rummel, Vincent Aleven. "Pair-Up: Prototyping Human-AI Co-orchestration of Dynamic Transitions between Individual and Collaborative Learning in the Classroom". Proceedings of the 2023 CHI conference on human factors in computing systems (CHI' 23, PDF)

Holstein & Olsen, 2021). At the intersection of education and HCI, there is an emerging emphasis on designing tools from a *human-AI co-orchestration* standpoint, which emphasizes the collaboration and shared control of orchestration responsibility between classroom stakeholders (usually instructor and students), and the AI system (Holstein, Aleven, et al., 2020; Holstein & Olsen, 2021; J. K. Olsen et al., 2020). The lens of co-orchestration is important: while assistance from the system with managing classroom activities is useful, instructors and students generally like to preserve some level of control over what happens (Echeverria et al., 2020; Holstein, Aleven, et al., 2020; Yang, Lawrence, et al., 2021). Adopting the human-AI co-orchestration lens, I (with the help of colleagues) created a technology ecosystem with the aim of allowing students to transition dynamically between individual and collaborative learning in a self-paced way. This technology ecosystem shares orchestration control with human instructors, in that AI systems recommend students' pairings to teachers, and teachers still preserve the control to make final orchestration decisions.

Dynamically combining individual and collaborative learning has pedagogical potential, but has yet to be investigated in authentic classrooms. In educational practice, combining individual and collaborative learning activities is very common (e.g., Think-Pair-Share (Kothiyal et al., 2014) and Jigsaw (Aronson, 1978)). From a learning science perspective, individual and collaborative learning each have their own benefits. From the point of view of the Knowledge Learning Instruction (KLI) framework (Koedinger et al., 2012), collaborative learning offers opportunities for mutual elaboration and co-construction of knowledge, whereas individual learning may promote induction and refinement as learning mechanisms (cf. (Koedinger et al., 2012; Mullins et al., 2011)). Previous work by Olsen et al. has shown that combining individual and collaborative learning may yield more effective learning for students than either mode alone (J. K. Olsen et al., 2019). Despite some emerging research on transitioning between individual and collaborative learning modes (Echeverria et al., 2020; J. K. Olsen et al., 2020), there has not yet been a study that empirically prototypes such processes in the classroom, supported by technology that teachers can use independently without much human assistance.

In this work, I conducted a classroom-based prototyping study of *dynamic transitions*. The main infrastructure is a co-orchestration technology ecosystem consisting of intelligent tutoring systems that support both individual and collaborative learning, and a novel teacher-facing co-orchestration tool named Pair-Up. This tool supports human-AI co-orchestration of students' dynamic transitions between individual and collaborative learning activities. It supports teacher awareness of students' learning status by displaying real-time learning analytics in both learning modes, and provides pairing suggestions to assist teachers with decision-making, but preserves space for teacher agency (i.e., teachers can flexibly override system-suggested pairings). I conducted a classroom prototyping study with five teachers and 199 middle-school students. The prototyping study aims to understand the feasibility and desirability of dynamically combining individual and collaborative learning in

K-12 classes, and get an initial sense of whether the human-AI co-orchestration technology ecosystem is up to the task. I performed mixed-methods analyses of user transaction log data collected from the technology ecosystem, as well as teacher interviews and student survey data.

5.2 Research Questions

Building on the idea of *fluid social transitions* (J. K. Olsen et al., 2020), I defined *dynamic transitions* as having three components: 1) monitoring students' learning skills, status and progress 2) having students transition when the need arises (e.g., when a student is no longer progressing productively in one mode of learning); and 3) pairing up students in ways that are not fully pre-planned. I developed a co-orchestration technology ecosystem that supports dynamic transitions and tested it in a classroom-based prototyping study. My study aims to answer the following research questions:

Observed classroom dynamics in classes with dynamic transitions between individual and collaborative learning

RQ1 What are students' learning processes?

RQ2 What are teachers' orchestration processes and pairing strategies?

Students' and teachers' perceptions of dynamic transitions between individual and collaborative learning

RQ3 Do teachers and students see value in dynamic transitions?

RQ4 What design opportunities do they see for future tools that might better support dynamic transitions?

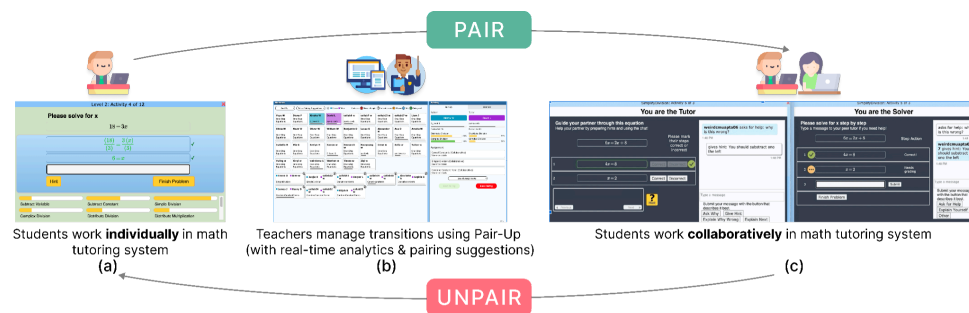


Figure 5.2. Co-Orchestration Technology Ecosystem for Dynamic Transition

5.3 Participants and Procedure of Classroom Prototyping

I conducted a prototyping study in authentic in-person classrooms in a suburban public school in the U.S. Five middle school math teachers (three males (T1, T4, T5) and two females (T2, T3)), and a total of 199 students from 11 classes voluntarily participated, including five 8th-grade classes, five 7th-grade classes, and one 6th-grade advanced class. The average class size is 19.5 (Min = 7, Max = 25). T3 teaches 7th-grade special education

with a smaller class size of 7 students, all of whom have an Individual Education Program (IEP). Most have a specific learning disability and may be 1-2 grade levels behind most 7th graders.

Each class participated for 2 class periods, resulting in a total of 22 study sessions. Each study session had a duration of 33-37 minutes so that each student participated for a total of 70-75 minutes. The first session of each class started with a 7-minute tutorial video that briefed the students about how to interact with the AI-based tutoring systems. All students started out in an individual learning mode of equation solving via their Chromebook with the individual tutoring system. A subset of the students (as assigned by teachers) collaborated using the collaborative tutoring system.

The teachers monitored and paired up students at any time during the class session, using the orchestration tool Pair-Up (Fig. 5.1) either through a desktop computer (T1, T2, T5) or a tablet (T1, T4). When pairing students, teachers were able to see students' data, including the skill bars and the number of problems finished. In addition to the data provided via Pair-Up, teachers were also walking around the room to keep an eye on students' pace and monitor their learning situation. The students who were paired up often did not sit next to their partner, and communicated with their partner via chat using the web-based collaborative tutoring system. When students who were paired up were done with the collaborative assignment, or when they were unpaired by the teacher, they switched back to individual work. After the classroom sessions, all teachers participated in a one-on-one semi-structured interview.

5.4 Data Collection and Analysis

To study the students' learning process (RQ1) and teachers' classroom orchestration processes (RQ2) during the dynamic transitions, I analyzed the log data collected from the two tutoring systems and Pair-Up (N = 19633 rows of transactions). Metrics related to students' learning process (Table 5.2) mainly include students' error rate, correct and incorrect attempts on each step, step duration, hint request frequency, and students' chat (in the collaborative tutoring system only). Metrics related to the teachers' orchestration process include the pairing algorithms that teachers selected, the students the system suggested to pair up, and the teachers' pairing actions. I now describe detailed procedures of the data analysis methods. To study how teachers and students perceive the value of dynamic transitions, as supported by the co-orchestration technology ecosystem (RQ3), and what design opportunities they see to better support such dynamic transitions (RQ4), I conducted and analyzed teacher interviews and student surveys.

Teacher Interview and Student Survey. The aim of the teacher interview was to understand whether teachers see value in dynamic transitions (RQ3) and what design opportunities teachers see for future tools to better support such dynamic transitions (RQ4). The interviews

lasted on average 45 minutes. The researcher leading the interview asked teachers about 1) the pairing strategy they used, 2) their experience with dynamic transitions, and 3) how supported they felt while using the tool. The teacher interviews were video-recorded and transcribed verbatim. Based on the interview transcripts of five teachers, two coders collaboratively conducted iterative affinity diagramming (Miles et al., 2013) for common themes to emerge. The teachers' statements were iteratively clustered into themes.

Table 5.1. Coding Scheme for Collaborative Learning Conversation Content Analysis

High-level categories	Definition	Sub-categories
Minimal Contribution	Behavior that involves little to no domain content knowledge, e.g., greeting, confirming partner's identity, chatting, or conversation related to usability and features of the tutoring systems	- Confirm partner identity - Minimal social behaviors - System usability-related conversation
Facilitative Contribution	Behaviors that involve domain content knowledge, and facilitate collaboration by moving the conversation forward, but are of limited help in building transferable skills for the domain knowledge	- A conversation that is generated by the tutoring software - Solver requests an answer - Tutor gives answers - Solver asks for correctness feedback - Tutor gives correct feedback - Other transactive social behavior
Constructive Contribution	A statement involving reasoning and explanation of content knowledge. For example, answering a question with an explanation, correcting others with an explanation, or asking a specific clarification question to help the partner build transferable skills	- Solvers ask Tutors for explanations/clarification - Tutors ask the solver for explanation/clarification - Solvers give explanations/clarification - Tutors give explanations/clarification

Additionally, I sent out a post-study survey to all participating students to understand whether students see value in dynamic transitions (RQ3) and study what design opportunities students see for future tools to better support such dynamic transitions (RQ4), and 171 students filled out the survey. The survey questions were adapted from a study by Lin et al. that evaluated students' perceived enjoyment and effectiveness in learning (Lin et al., 2020). The survey contains seven five-point Likert scale questions about students' *enjoyment and perceived effectiveness* in learning individually and collaboratively (where 1 means "strongly disagree" and 5 "strongly agree"), two multiple choice questions on *students' preferred level of agency* in the pairing process, and three open-ended questions about: 1) how students like the experience of switching between individual and collaborative learning, and what they 2) like or 3) not like about the experience. I analyzed students' responses to the multiple-choice

questions by computing the percentage of students who chose each option, and their responses to the Likert scale questions by computing means and standard deviations. To test whether students perceived the effectiveness and enjoyment of individual and collaborative learning modes differently, I performed paired t-tests on students' responses to Likert scale questions. Students' responses to the three open-ended questions were coded by two coders through thematic analysis (Braun & Clarke, 2016) to identify common themes in students' opinions.

Collaborative Learning Conversation Content. To understand students' learning processes during dynamic transitions (RQ1) and probe deeper into students' collaborative learning processes, I performed a manual content analysis of students' conversations. In this study, students communicate with each other via the chat window. Similar to the method used in analyzing students' chat data in peer learning (cf. (Mawasi et al., 2020; X. Wang et al., 2021)), I extracted chat data from students' conversations, logged by the collaborative tutoring software. I adopted the coding scheme from Mawasi et al. (Mawasi et al., 2020), where they coded peer help-giving behaviors in middle school math classrooms, with three categories for student conversation: minimal contribution, facilitative contribution, and constructive contribution, under which I developed 13 subcategories (Table 5.1).

Analyzing Teachers' Pairing Strategies. To understand teachers' orchestration progress and strategies for pairing up students (RQ2), I extracted from log data the number of times teachers pair and unpair students, the pairing algorithms used, and whether they followed system suggestions. Additionally, I investigated whether students who were assigned to different collaborative roles by their teacher exhibited different patterns in their learning. Specifically, I compared the learning curves (cf. (Cen et al., 2006)) of students who were assigned to take on different roles (Solver, Tutor, both, or neither). One motivation behind this analysis is to confirm whether teachers' actual pairing selection aligned with their intended pairing strategy. Learning curves are auto-generated line graphs based on the transaction data students' interaction with the tutoring system in Datashop⁸. My analysis procedure contains the following steps:

- *Step 1.* Depending on the role teachers assigned to students when they paired them up, I grouped students into four categories: students who only served in the role of Tutor (Tutor Only, $N = 53$), students who only had the role of Solver (Solver Only, $N = 61$), students who had both roles, though at different times (Both Solver and Tutor, $N = 64$), and students who did not do any collaboration (No Collaboration, $N = 20$).
- *Step 2.* I created four data subsets, one for each of the four categories defined in Step 1. I emphasize that these groups were formed after the classroom sessions in the

⁸ A data analysis repository and analytics infrastructure for the learning sciences community (<https://pslcdatashop.web.cmu.edu/>)

analysis process based on teachers' decisions. Thus, the groups are not experimental conditions with randomly assigned students.

I visualized the learning curve of each subset, and compared the following metrics for each subset: Error Rate, Number of Incorrect (Steps), Number of Hints, Step Duration, Correct Step Duration, Error Step Duration (Table 5.2). Below, I report where I find salient differences in terms of the learning curve (i.e., indicated by one line of a group consistently higher or lower than another comparison group).

Table 5.2. Relevant Metrics on Students' Learning From Transaction Data Logged by Intelligent Tutoring Systems

Measure	Description	
Error Rate	The percentage of students who asked for a hint or were incorrect on their first attempt.	
Assistance Score	For a given opportunity, the number of incorrect attempts plus hint requests equals the assistance score.	
Number of Incorrects	of	Total number of incorrect attempts by the student on the step
Number of hints	Total number of hints requested by the student for the step.	
Step Duration	Total time spent on a step, calculated by adding all of the durations for transactions that were attributed to a given step.	
Correct Step Duration	Step	The step duration if the first attempt for the step was correct. Might also be described as "reaction time" since it's the duration of time from the previous transaction or problem start event to the first correct attempt.
Error Step Duration	The step duration if the first attempt for the step was an error (an incorrect attempt or hint request).	

5.5 Results: Students' Learning and Teachers' Orchestration Process During Dynamic Transitions (RQ1 and RQ2)

In this section, I report students' learning processes that occurred as they switched back and forth between individual and collaborative learning (RQ1), based on log data collected by the tutoring systems and Pair-Up. As an overview of the study process, I found that students solved, on average, 23.8 math problems individually and 1.7 problems collaboratively. Students spent on average 39.8 minutes (76% of total time) on individual problems and 12.0 minutes (24% of the time) on collaborative problems (Table 5.3). As shown in Table 5.3, during the 22 class sessions, 210 collaboration activities (defined as two students teamed up to work collaboratively on one assignment) happened, with an average of 18 activities in each class. This considerable number of collaboration activities is an initial piece of evidence that

teachers were well-supported in orchestrating dynamic transitions by the co-orchestration technology ecosystem.

Table 5.3. Analysis Results From Log Data Collected in Tutoring Software and the Orchestration Tool (n = 19,633 Rows)

Analysis	Mean	SD	Min	Max
Number of collaboration activities per class (over 2 class sessions)	18.0	8.5	6	37
Math problems solved individually per student	23.8 (93%)	17.2	0	70
Math problems solved collaboratively per student	1.7 (7%)	2.0	0	7
Time spent on individual problems per student (in min)	39.8 (76%)	14.5	0	80
Time spent on collaborative problem per student (in min)	12 (24%)	7.9	0	32.8
Number of unpairing actions by the teacher per session	2.8 (22%)	2.5	0	10
Number of times teachers followed the pairing suggestions	6.4 (48%)	4.9	1	20

Students' Collaborative Learning Processes (RQ1)

In this section, I report results on students' observed learning processes as they engaged in dynamic transitions (RQ1), based on collaborative learning conversation content analysis.

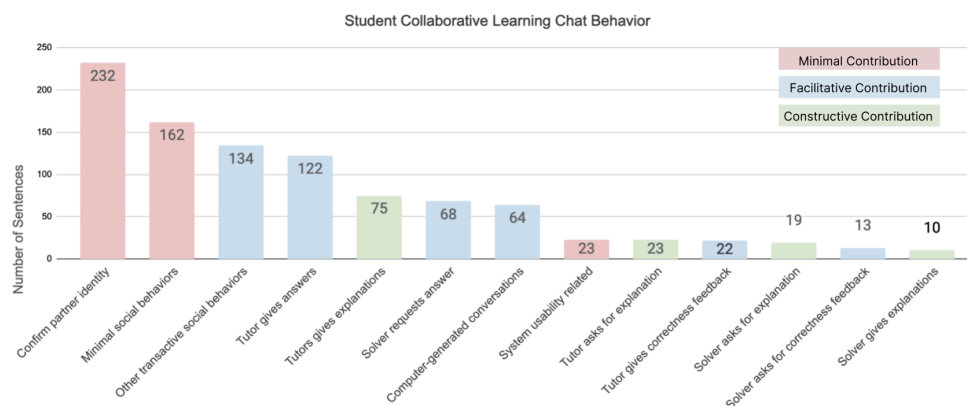


Figure 5.3. Coding Results of Students' Chat Behaviors During Collaborative Learning

Table 5.4. Representative Examples of Student Conversation for Minimal, Facilitative, or Constructive Contribution (Students' Names Are Aliases)

Minimal	Facilitative	Constructive
T: who are you	S: hello blozo	S: it probably wrong
S: Ash who are you	T: Who are you	T: it is a negative number
T: Angela	S: enzo	T: would you like me to explain the steps of the problem?
S: OKay	T: Wrong	S: yeah plz im confused
T: soooo am i supposed to like tell u what to do	S: why tho	T: alright! just gimme a second
S: yeah	T: You need a =	T: so first, you'd subtract the 5 from the 1 so everything is on either side [Tutor gives explanations]
T: Ok	S: bruh how it wrong	T: let me know what you get when you subtract 5 from 1
T: oh wait no ur supposed to tell me your steps and i tell u if its right	T: Idk	S: -4
	T: type $2*x+2*-3+6=14$	T: okay, so you take that -4 and divide it by 4
	[Tutor gives answers]	T: what do you get when you divided -4 by 4? [Tutor asks for explanation]
	T: put that as your answer..thats what i was told to tell you	S: so it would be -1
		T: yes :)
		T: $x=-1$

As mentioned, I analyzed the content of students' dialogues using the coding scheme shown in Table 5.1 to see what kind of conversational behavior students engaged in during their collaboration. Of all coded behaviors in collaborative learning conversations, 43.1% were minimal contributions, 43.7% were facilitative contributions, and 13.1% were constructive contributions. The standard deviation of the code frequencies tended to be large, suggesting that different student pairs can exhibit different learning behaviors. The top subcategories within each major category are: Confirming partner identity (Minimal), transactive social behaviors (Facilitative), and Tutor gives explanation/ clarification (Constructive). The results from this analysis (details in Fig. 5.3) show that students engaged mostly in minimal conversational behaviors such as building rapport or confirming their partners' identity, as well as in facilitative behavior such as giving answers directly to their partners without scaffolding. They engaged less frequently in constructive behaviors that may deal with conceptual knowledge and hold the potential to build transferable mathematics skills. Typical conversations are shown in Table 5.4.

Teachers' Orchestration Processes and Pairing Strategies (RQ2)

In this section, I report results on how teachers orchestrate the dynamic transitions and on their strategies for pairing up students (RQ2), based on results from the teacher interviews, log data collected from Pair-Up, and the analysis of teachers' pairing strategies.

From classroom observation, I found that two teachers (T2, T3) teamed up all students at the same time, and three (T1, T4, T5) paired students up at different times. From the interview data, I ascertained that T3 (who taught special education students) thinks pairing and unpairing everyone at the same time would make it easier for her students (who usually have some kind of learning disability) to manage the transitions, and easier for her to manage the classroom since that way she does not need to give instructions multiple times. Although I didn't have a chance to ask T2 why she paired up everyone at the same time, I speculate that

she prefers more synchronized classroom management because she taught the largest number of classes (five sessions a day) and students (104 in total).

It is an interesting question to ask about what information teachers pay attention to as they make decisions about pairing up students. One teacher mentioned that they would read the room: *“I mean, uh, some classes, like some of my periods, they don't require a lot from me. Mm-hmm, they don't need as much guidance or assistance or help. And so then I would just kind of monitor and walk around and see, or watch on the screen to see when I should pair or when it looks like they're ready to be paired or ready to switch roles”* (T2). Another teacher mentioned that they would look at students' skill bar and progress on Pair-Up, before checking system pairing suggestions: *“I was looking for kids that were not doing so well. And just natural instinct, right away, I wanted to pair them with my dashboard; somebody had their bars up high. I would feel safer pairing when a student has someone as a tutor who has more knowledge, at least on the bars there. Then I started looking at the suggestions from the tutor themselves”* (T1).

I also looked at the degree to which teachers followed the suggestions given by Pair-Up regarding which students to team up. As mentioned, the Pair-Up tool gave pairing suggestions under two pairing algorithms, and teachers had the final say about whether to follow, ignore, or override these suggestions. Analysis of the log data from the Pair-Up tool (Table 5.3) indicates that teachers followed the tool's pairing suggestions 6.4 times per class session, comprising 48% of their total pairings. Teachers followed the suggestions of the Different Knowledge Level algorithm (82%) more frequently than the Random algorithm (18%). There was substantial variability between teachers in the degree to which they followed the tool's pairing suggestions, ranging from 23% of all pairing decisions to 86%. Three teachers (T1, T2, T3) found pairing suggestions to be helpful *“some of the suggestions were good suggestions, I think.”* (T2), and *“I also liked the one that, um, [the system] gave the options for who to pair up, like the suggestions, I like that”* (T3); while one teacher (T4) preferred to manually pair up students, as he thought he was more familiar with students' situations and can pair students in a more flexible and personalized way: *“I didn't really use any of the pairing suggestions for that class. Just because I've been working with them for quite a while, I kind of already know the students. I go more with just the manual because I take all that other stuff into consideration too when I'm pairing them up. I like to kind of leave it open and just kind of look and see who needs paired up and who needs help with what, rather than worry about suggestions”* (T4). Teachers may reject system's suggestions, e.g., if the students suggested don't get along (*“The only one I was considering changing was Tom and Roy (alias), just because of, again, social interactions that they've had. They don't get along”*) (T3), or if teachers want to see how different students interact: *“Sometimes I override it just to see how certain students would work together; to see how these students would interact with each other”* (T1).

Based on the teacher interviews, I found that teachers generally wished to pair students doing well with those not doing well (i.e., struggling). For example, teachers mentioned that *“I try to pair the students who were really good at a topic with someone who may be made some mistakes on that topic before, I uses the students who know farther ahead and obviously had already mastered a lot of those topics and skills as the tutor”* (T4) and *“..if you see somebody who, you know, has really got it and then somebody who’s maybe struggling on the problems a little bit, then I pair them up”* (T3). Teachers also mentioned taking into account students’ interpersonal relationships: *“A lot of them who are paired are friends. They are more familiar with each other. They can be more direct in their evaluation of what the kid is doing”* (T5). Pairing up friends may ensure they are comfortable working together and increase interaction between collaboration partners. *“Especially in a short period of time. I know I just had two periods. If this is over the course of a school year, I may mix them all up. I just knew I had limited time, and I wanted to get as much interaction as I could. I’ve seen this where there’s not much communicating going on. It’s just two individuals sitting beside each other sort of doing their own thing. You know, the collaboration is good, but the kids have to buy into it.”* Teachers also said they take into account personalities and physical proximity in class. These intended pairing criteria (as stated in interviews) generally aligned with prior research (Yang, Echeverria, et al., 2021b; Yang, Lawrence, et al., 2021).

As for the timing of when teachers paired up students and the conditions under which they do so, some teachers (T3, T4) paired up students based on their observation and monitoring of students’ progress in the individual activities, and pair up students who are continually struggling (*“As students have progressed a little bit, I basically look at how they’re progressing and if I see a students continually struggling, then that’s where I would be, pairing them up”* (T4)). For the teachers who paired up students dynamically at different times, they either saw the timing of pairing as something that can be flexibly adapted to students’ unique situation (*“But I pair them up when I feel they need it. There is not a set time that I do it”* (T5)), or they view the timing as not critical (*“I don’t see a bad time”* (T1)).

Based on the log data, I find that on average, teachers unpaired students 2.8 times per session, representing 22% of all the pairs they formed. Other unpairing happened automatically when students finished their collaborative assignments. As to the timing or conditions when teachers unpair students, the five teachers in the study had different preferences. Two teachers (T1, T4) liked to unpair students when their knowledge level increases (*“[I unpair students] whenever they have started to start to show mastery of a new concept if they were paired up for that reason”* (T1)), when their “misconceptions are gone”, or if they needed no additional help (*“If I saw the student was able to get through the first two questions with the help and they didn’t need any additional help, then I wouldn’t pair them and let them go back to individual”* (T4)). One teacher (T5) said they would unpair students based on how the collaborative activity is progressing, for example, when it becomes social time (*“You hear the conversation, and then the conversation is not focused on what they’re doing anymore. And it just becomes, you know, then just chit chatting, then that’s when I know we gotta get away*

from this” (T5). This teacher also stated “...you can’t do it too much. because I think too much collaboration,...it turns into social time. It’s not productive, there’s a fine line” (T5). Some teachers mentioned they didn’t want to limit the collaboration to a fixed number of problems, or “haven’t really thought honestly about what would be the appropriate time to unpair them if we were doing something (learning activities) after” (T4).

Regarding the question of whether teachers’ actual pairing decisions aligned with their intended pairing strategy, the quantitative analysis of the teachers’ pairing strategies produces three relevant findings:

1) The No Collaboration subgroup performed worse than the other subgroups. After comparing the No Collaboration group to the other three groups combined with regard to all metrics listed in Table 5.2, I found students who were not selected by their teacher for collaborative assignments (i.e., the No Collaboration group, who only worked individually) generally made more mistakes (in their work on the individual tutoring system) than the students in the other groups, all of whom were involved in collaboration (the No Collaboration group has the highest error rate across learning opportunities). I also found from the learning curve that the No Collaboration students spent more time per step than the other three groups not only on incorrect steps but also on correct ones.

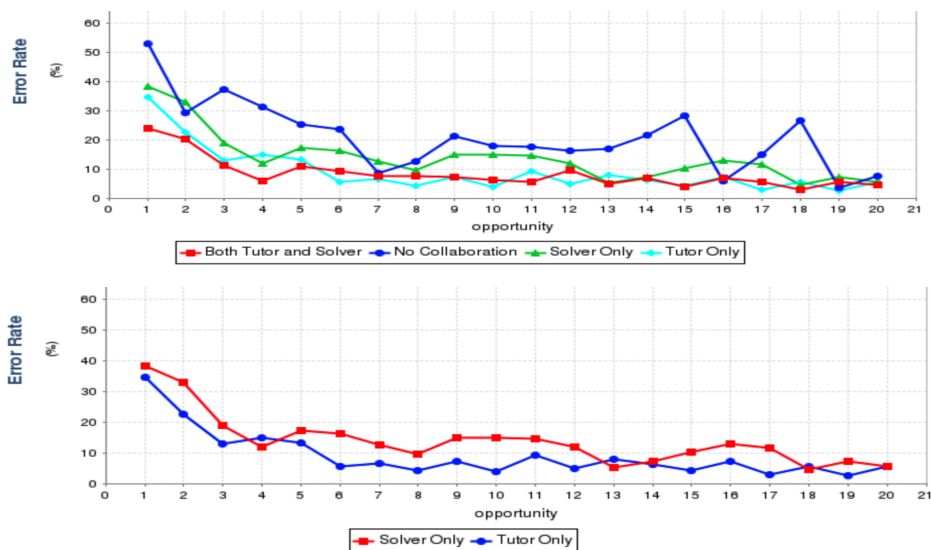


Figure 5.4. Learning Curves of Error Rate: No-Collaboration Group vs. Other Three Groups (Left); Solver-Only vs. Tutor-Only (Right). The X-Axis (Opportunity Count) Refers to the Number of Practice Opportunities Students Had With a Given Knowledge Component in the Domain Content

I conducted statistical tests comparing the error rate of No Collaboration subgroup with other subgroups, and the error rate of No Collaboration group was significantly different as compared with the three subgroups combined ($M = 0.23$ vs. 0.15 , $SD = 0.18$ vs. 0.16), $t(194) = 2.06$, $p = 0.02$. Similarly, I conducted t-test which shows, No Collaboration group spent

significantly longer time duration (in seconds) than the other three groups ($M = 34.26$ vs. 22.78 , $SD = 20.06$ vs. 28.57), $t(194)=1.69$, $p = 0.04$.

This evidence suggests that the students who were left to work individually by their teachers could use help from an instructor or from peers. It may indicate that there is room for improvement in the co-orchestration process, in that it could better help teachers to identify students who might benefit from being paired up. While I cannot know the specific rationale teachers have for leaving each student without collaboration, I can infer, from interview, three factors that may have played a role: 1) student behavioral: the teacher expects that certain students might not work well with others, 2) student preference: the teacher thinks the certain students might prefer learning by themselves, (e.g., “*Um, I have a few, I have some students that do not work well with other students and want work on their own and [I like] just having that opportunity for them to work on the solo assignment.*” (T2)) or 3) the teacher has helped them individually (e.g., “[*I would*] first provide individual help, then the pairing serves as the additional help” T4).

2) Tutors Only performed somewhat better than Solvers Only. Students who were selected to take on the role of Tutor but not the role of Solver (i.e., the Tutor Only group) had better performance on the individual learning activities than those who only took on the role of Solver. The learning curves derived from the log data from the tutoring system suggest that the *Tutor Only* group had a lower error rate than the *Solver Only* group. There were no obvious visual differences in the learning curve on step duration for the Tutor Only and Solver Only groups. Neither the difference between the Tutor Only group and the Solver Only group in error rate ($M = 35.61$ vs. 22.65 , $SD = 20.06$ vs. 28.57), $t(110)=0.23$, $p = 0.41$), nor that in step duration ($M = 27.73$ vs. 22.5 , $SD = 28.80$ vs. 40.45), $t(110)=0.80$, $p = 0.21$) was statistically significant. There was some evidence that teachers preferred (and were able to select) higher-performing students in the role of tutors, though the performance difference between students selected to only be tutors v. those selected to only be solvers was not statistically significant. They may view these students as more capable of helping students with weaker learning or slower progress. This finding aligns with most teachers’ intended pairing strategy as stated during the interviews (i.e., to pair those who learn well with those who are learning less effectively), as well as the most common strategy teachers generally adopted as found in prior research (Yang, Lawrence, et al., 2021).

5.6 Results: Teachers’ and Students’ Perceived Value and Design Opportunities (RQ3 and RQ4)

This section reports results regarding teachers’ and students’ perceived value of dynamic transitions (RQ3), as well as the opportunities for future tool design expressed by them (RQ4), based on data from the teacher interviews and student surveys.

Teachers' Perceived Value of Dynamic Transitions (RQ3)

In this section, I report interview results on teachers' experience with and perceived value of dynamic transitions (RQ3). All five teachers (including the special education teacher) reported being likely to use technology such as the Pair-Up ecosystem in their regular classrooms.

When asked in interviews, teachers generally saw pedagogical value in dynamically combining individual and collaborative learning activities. As one teacher expressed: *"I definitely like it (the dynamic transitions) because it lets you differentiate the learning a little bit better because some kids don't need to be paired up"* (T4), and *"I do like that aspect of having it where you can pair up some or have some work individually, and you can mix it up as needed"* (T4). One teacher liked that this self-paced learning can help distribute resources, support peer learning, and allow students in a class to catch up: *"You got to separate the kids, kids who all understand everything, some students in the middle. Got to use those who get it very quick to try to help the others"* (T5). Another teacher (T2) thought it was useful *"to let students who do not work well with other students and want to work on their own, just have the opportunity to work on solo assignments"*. She also stated that dynamic transitions are useful for students who are *"at home due to COVID for an extended period of time, and might not be logging on at the same time"* (T2). However, the special education teacher (T3) had some reservations about dynamic transitions for her class. Though she thought her students liked a mix of different activities and *"enjoyed (switching) back and forth instead of just doing the same thing the same way"*, she considered that dynamic transitions are useful only *"for students who can handle the transitions."* She was concerned that some of her students with learning disabilities may find the transitions back and forth between different types of activities to be difficult. She still found the dynamic transitions worthwhile and mentioned that *"students learn better when they switch (between learning modes), so I think the pros outweigh the cons"* (T3). One teacher thought that switching back and forth too frequently may hinder students' learning, as they may get distracted about who their partners are (T5). Teachers also think that more advanced classes may manage dynamic transitions better. For example, one teacher (T4) selected a smarter class for study participation as he thought the students who had less experience with equation solving would struggle with the math content knowledge and might not have the headroom to simultaneously take on the challenge of navigating smoothly between the two tutoring systems.

As for teachers' perceived classroom management load compared to the regular load, the two teachers who used the orchestration tool on *tablet computers* mentioned it was easier or similar to their regular classes. One teacher (T4) expressed that it was easier because he can multitask (walking around the classroom while pairing students), which saves them time so they have more time to help individual students. *"Because it was easier to work with more students and see what they were doing and I can do a quick pair up instead of maybe taking five minutes like I would have to normal class to help certain students"*. This teacher thinks it

would be harder if he had to use a static device such as a desktop computer “*I think it enhances that if it's mobile. I think it would totally take (me) away from the classroom management if I have to sit at the desk*” (T4). The special education teacher (who used a desktop computer) described it was harder to manage the classroom.

Overall, teachers mentioned they liked the experience, for four main reasons: 1) it allows a possibility for self-paced learning personalized to students’ learning progress and preferences, 2) it allowed teachers who use the orchestration tool on the mobile device to multitask (e.g., pairing, monitoring and helping students), 3) it provided a different way of teaching compared with their regular lecturing, 4) teachers like that the software provided support for peer tutoring, in addition for support for math learning.

Students’ Perceived Value of Dynamic Transitions (RQ3)

In this section, I report quantitative and qualitative survey results on students’ experience with and perceived value of dynamic transitions (RQ3). From the 171 responses of *five-point Likert Scale questions* in students’ survey (Table 5.5), over half of the students rated a mix of individual and collaborative learning as helpful for their math learning (54%). 30% were neutral and 16% of students disagreed to a greater or lesser degree ($M=3.3$ out of 5, $SD=1.05$). Students reported higher enjoyment in individual learning than collaborative learning, and the difference was statistically significant ($M=3.5$ v.s. 3.3, $SD = 1.16$ vs. 1.23, $t(169) = 2.18$, $p=0.03$). 60% of students liked working individually using the online math tutor, and 49% of students liked working collaboratively (chose “like somewhat” or “like a great deal”). Students also perceived *individual learning to be more effective* and the difference was not statistically significant ($M=2.9$ v.s. 2.7, $SD=0.99$ v.s. 1.06), $t(169) = 1.36$, $p=0.08$. More students preferred working individually all the time than working collaboratively all the time ($M=2.8$ v.s. 2.6, $SD=1.27$ v.s. 1.23), and the difference was statistically significant, $t(169) = 2.06$, $p = 0.04$.

The three optional *open-ended* questions (1) how students like the experience of switching between individual and collaborative learning, and what they 2) like or 3) do not like about the experience), respectively yielded 48, 50, and 47 students’ responses. Based on these responses, I found that 39% of students liked switching between individual and collaborative learning ($N=19$). Reasons that students like dynamically combining individual and collaborative learning include: “*It was helpful so you could do things on your own but also be able to have someone to be partnered with,*” or “*It was very cool and different from other sites like this.*” 31% of students rated the experience as neutral ($N=15$), and 25% reported not liking it ($N=12$). Analyzing why these 12 students did not like the switching between activities, it turned out to be mainly because they preferred a single learning mode ($N=7$), e.g., “*I don’t like the switch cause I was doing good on my own*”. While most students are fine with switching to a collaborative activity while solving a problem individually, a minority of students reported not liking the unanticipated switching ($N=2$) e.g., “*It was fine,*

a little annoying at times because I would be in the middle of a question (and I had to leave it)”, indicating design opportunities for the transitions to happen in an even smoother and less intrusive way.

Students enjoyed collaborative learning with dynamic transitions to different degrees. Asked what they enjoyed about the study experience, among the 50 students’ responses, 40% of students (N=20) reported enjoying collaborative learning and having a partner to work with, and liked *“working with random people from across the room”* or *“working with a friend”*. Only 6% expressed that the thing they enjoy is working alone (N=3). When asked the one thing they don’t like about the experience, among the 47 responses, 30% of students mentioned some aspects of the collaborative learning experience they didn’t enjoy (N=14). The main complaints students have about the collaborative learning experience include their partner taking too long to answer or not doing the work (N=5), students not liking their assigned tutor (N=4), or not being able to pick their partners (N=4).

Table 5.5. Survey Responses From 171 Students on the Experience of Dynamic Transitions

Survey Items (Self-report by students in five-point Likert Scale)	M	SD
How much did you like working individually using the online math tutor?	3.5	1.16
How effective did you learn math knowledge through individually using the online math tutor?	2.9	0.99
Would you prefer to work individually all the time?	2.8	1.27
How much did you like working collaboratively using the online math tutor?	3.3	1.24
How effective did you learn math knowledge through collaboratively using the online math tutor?	2.7	1.06
Would you prefer to work collaboratively all the time?	2.6	1.23
Do you think a mix of individual and collaborative learning is helpful for your math learning?	3.5	1.04

Design Opportunities for Future Tools Grounded in Data (RQ4)

In this section, I report design opportunities for future tool design (RQ4), based on results from both students’ surveys and teacher interviews.

Students’ Opinions on Student Control and Agency in the Dynamic Transitions

The survey results indicate that many students prefer to know in advance when and with whom they will get paired. Only 14% of students reported they are okay with not knowing the information. 37% of the students said they would like to know *both* when and with whom they will be paired. Students cared slightly more about with whom they are paired (27%) than exactly when they will be paired (22%). In addition, the majority of the students (59%) *wished they could have chosen their partner*, 33% reported “Maybe”, and 8% reported “No”. These findings suggest that co-orchestration tools for transitioning between learning activities may need to support greater awareness on the part of students and agency regarding the choice of the partner and the timing of transitions, which could be improved in future iterations of the tools.

Teachers' Opinions on Student Control and Agency in the Dynamic Transitions

One teacher mentioned he would like to give students some control over when to collaborate, but not full control, as some students “*will just mess around. (T1)*”. Most teachers preferred that students not know who their partner is (N=3). Results from affinity diagramming reveal that teachers thought anonymous pairing may 1) avoid some social issues (e.g., students do not get along with their partner), 2) avoid some hard-to-control student behaviors that make classroom management harder (e.g., shouting across the room), 3) prevent students from off-topic chatting, 4) protect students' privacy, and 5) in the special education class, prevent students from making fun of smart students and teasing them about being the “Tutor”. One teacher, however, acknowledged that it is hard to prevent students from knowing their partners' identities.

Teachers' General Design Feedback

Teachers made several suggestions for improving the tool. 1) Currently, the tool suggests several possible partners for any given Solver. Two teachers expressed a preference for even more efficient pairing, with the tool automatically pairing up students and the teacher only having to review and confirm these pairings. 2) Currently, the roles are fixed for any collaborative activity. One teacher desired a simple way to have students do reciprocal peer tutoring, to experience being both a Solver and a Tutor, which was the motivation for the first version of the collaborative tutoring system (Walker et al., 2014). 3) Currently, the system assigns students to a new collaborative problem when they are paired. One teacher strongly preferred to just pair students to work collaboratively on the problem they were initially working (struggling) on individually, because it would be more effective to help struggling students who may just need “a little additional nudge or help” to get unstuck.

Two teachers (T1, T4) wanted to be able to see detailed, deep-dive information about individual students (e.g., (Holstein, McLaren, et al., 2018)), so that they could provide personalized help and pair students in a more customized way. One of these teachers prefers system suggestions on what collaboration activity students should be assigned based on students' misconceptions. The special education teacher made slightly different suggestions than the other teachers. Her students may easily lose focus and are very “reward-motivated”. Thus, she suggested adding visual or audio rewards (e.g., a “Ding” sound) when they did something right to keep them engaged and make the transition between activities even smoother. She would also appreciate more tips and guidance on when to pair or unpair students.

5.7 Learning and Orchestration Process of Dynamic Transitions, and New Classroom Dynamics

I found students in collaboration activities in dynamic transitions engaged mostly in minimal conversational behaviors such as building rapport or confirming their partners' identity, as

well as in facilitative behavior such as giving answers directly to their partners without scaffolding. They engaged less frequently in constructive behaviors that may deal with conceptual knowledge and hold the potential to build transferable mathematics skills (RQ1). I found some evidence that teachers preferred and were able to identify higher-performing students in the role of Tutors, so as they facilitate their partner's learning. However, I also found that students who kept working individually (i.e., who were not assigned to collaborative learning by their teacher) performed worse than those who were assigned to collaborate (RQ2). Potential reasons that teachers did not select students for collaborative learning activities include behavioral reasons (certain students might not work well with others), student preference (the teacher thinks the students might prefer learning by themselves), or the teacher would like to help them individually. Still, assuming peer tutoring can be an effective alternative in getting students unstuck, leaving students in individual mode may risk prolonging or exacerbating students' struggles, especially if teacher help is not adequate to help them. Thus, it may be fruitful for future orchestration tools to aim to provide even more effective support in guiding teachers in identifying students who need to be paired up.

With this initial exploration of dynamic transitions in the classroom, *new classroom dynamics* arose. Firstly, the students switched dynamically between (AI-supported) individual learning and (AI-supported) collaborative learning. Secondly, teachers were able to successfully orchestrate this complex process and multitasked. Teachers walked around the classroom to keep an eye on individual students while monitoring the progress of all students and quickly pair students up using the orchestration tool, accepting the tool's suggestions as to whom to team up or substituting other choices. Designing orchestration tools that allow teachers to multitask has been explored (cf. (Kharrufa et al., 2017)), indicating an exciting venue for design exploration for future tools, as it fits the practical needs in teachers' classroom practice (An et al., 2020; Holstein et al., 2019b). Thirdly, teachers expressed that dynamic transitions could be helpful given COVID-19 constraints: as students may join remotely or log in at different times, it allows them to be paired up and collaborate more flexibly without prolonging the waiting time for partners.

Overall, the prototyping study provides evidence that, with appropriate technology support, dynamic combinations of individual and collaborative learning are feasible - the orchestration is manageable, teachers appear to be content with the teams they form, which tended to align with their stated preference to assign higher-performing students to the role of Tutor. Students were able to switch between the two tutoring systems without great problems - even if there were some relatively minor complaints regarding the suddenness of transitions and less agency over the choice of partners.

5.8 Experiences with and Perceived Value of Dynamic Transitions

For RQ3, I found that students' perception of dynamic transitions can be affected by their collaborative learning experience. A considerable number of students report certain aspects of collaborative learning (usually concerning the partner they are paired with), to be the one thing they particularly like (40%) or dislike (30%) during the entire study experience. This indicates that collaboration can be a more polarizing experience, which may be inseparably tied to and greatly reflect their perceived enjoyment and effectiveness of *working with a particular partner*, instead of collaborative learning activities themselves. This finding may also explain why students desire more agency in the pairing process (i.e., knowing when and with whom they are paired, and choosing their own partner).

From the teachers' perspective, I found initial evidence that teachers felt well supported by the Pair-Up tool in orchestrating dynamic transitions. As well, they liked being able to monitor students and pair them up based on their progress in individual activities. Teachers reported a positive view towards dynamic transitions and expressed that they were likely to use the technology in their regular teaching in the future, as it allows for more differentiated learning and it helps ensure that students receive help in moments of struggle. Nevertheless, my findings also turned up some concerns from special education teachers or teachers of weaker classes, namely, that students with weaker domain knowledge or learning disabilities may not handle switching between activities well.

Most teachers (four out of five) suggested that the classroom management load when using the technology ecosystem is similar to or less than that in their regular practice. In general, teachers were receptive to the idea of dynamic transitions and recognized the pedagogical value of dynamically transitioning students from individual to collaborative learning and back. This finding confirms results from a prior co-design study (Lawrence et al., 2022). Over half of the students thought a mix of individual and collaborative learning is helpful for their math learning. Prior research shows individual learning may yield higher learning gains for *procedural knowledge* (J. K. Olsen et al., 2015), which arguably falls under this context (math algebra), thus it is not surprising that students perceive individual learning to be more enjoyable and effective than collaborative learning.

While it was not the primary focus of this study to evaluate which type of device affects the orchestration load of dynamic transitions, teachers perceived that a mobile device (e.g., a tablet) yields a lower orchestration load compared to a static device (e.g., a desktop) when orchestrating dynamic transitions, as it allows them to multitask in their usual practice. Prior research in the HCI and AIED communities has demonstrated that mobile, portable, or wearable devices in classrooms hold great potential to support more efficient teaching practices and classroom orchestration (c.f. Keeping watch (Quintana et al., 2016), ClassBeacon (An et al., 2019), Lumilo (Holstein et al., 2019a), and others (An et al., 2020;

Bernacki et al., 2020; Montrieux et al., 2015)), which this work provides additional evidence for.

5.9 Design Implications and Challenges

For RQ4, in addition to the specific suggestions pointed out by teachers and students in surveys and interviews, I provide the following *design implications* based on study findings, intended for the HCI and educational research communities, especially researchers building human-in-the-loop technologies to support teaching and learning.

Firstly, future tools for supporting dynamic transitions or other chat-based short-term collaboration activities should focus on cultivating positive and constructive collaboration behavior (which involves reasoning and explanation of content knowledge and may help build transferable skills). Constructive contributions in collaborative dialogues, such as reasoning or explanation in terms of conceptual knowledge, can often support students in building transferable skills (Bartel et al., n.d.; Mawasi et al., 2020). In this study, constructive contributions were rare in students' dialogues, however. Scaffolding students in better peer-tutoring, so that they engage in more constructive behaviors such as explaining and clarifying conceptual knowledge, may lead to more productive collaborative learning experiences. In such short-term collaboration activities, it can also be important to ensure that the students are responsive to their partners. I envision one way of nudging students for more timely interaction is for the system to display a timer countdown notification when one party is idle for a long time.

Secondly, teachers would like students' analytics that more directly inform teachers of the optimal timing of pairing and unpairing actions. While students are working *individually*, teachers desire more in-depth information on students' misconceptions and current problems to help them make more informed pairing decisions. In *collaboration*, given the considerably different collaboration experiences and large variance in collaborative learning quality, it can be important to help teachers monitor the collaboration progress and quality, so as to maximally optimize the learning experience and outcome for students. Such a design for collaborative learning analytics can draw from prior work on positive indications of collaboration, including sharing resources, joint actions, mutual planning, equal participation, communication, and reaching consensus (Kharrufa et al., 2017). Additionally, one concrete learning analytics that might guide teachers in searching for optimal moments for unpairing students can be for the tool to conceivably notify teachers when students have improved substantially on the knowledge skills that they were paired up for initially, or when their misconceptions were lifted.

Thirdly, tools for dynamic transitions may need to be *customizable and support teachers' varied preferences depending on their goals for pairing*. Educational technology works best when aligned with teachers' values, instructional goals, and teaching practice (Holstein et al.,

2017; Kirschner, 2015). Echoing the technology probe study by Echeverria et al., who found that the design of dynamic transitions may need to be adaptable for different classroom contexts (e.g., teacher's preference and students' prior knowledge) (Echeverria et al., 2020), the current study uncovered that the tool design may need to be adaptable to the *goals, purposes and values* teachers hold for initiating such collaboration. For example, for teachers whose goal for pairing students is *to have reciprocal peer-tutoring*, they prefer a role-switching feature, so students exchange the roles of Tutor and Solver. In contrast, teachers whose goal for pairing is *to help struggling students get unstuck* wished for a feature to allow struggling students to continue working on the problems they were stuck on in collaboration mode with a peer partner's help, instead of getting new problems. Some other teachers, who teach large classes, may prefer a more efficient and automated way of forming pairs. They may be more willing to sacrifice agency and control in the human-AI co-orchestration process in exchange for greater efficiency (Amershi et al., 2019).

This study also reveals a *design challenge*: I observed a tension between teachers and students' preferred level of control during the transition process. Previous studies in human-AI co-orchestration found that teachers prefer to have the final say in the pairing process (Echeverria et al., 2020; Yang, Lawrence, et al., 2021). However, in my study, we found that students would also like to have control over when and with whom to be paired. How to creatively incorporate more student agency without threatening teachers' authority and preferred control of the classroom presents a design challenge. As Holstein and Olsen pointed out, sharing the orchestration responsibilities between instructors and learners may be beneficial, but may also risk creating a greater load for the instructor (Holstein & Olsen, 2021). Still, it is an emerging vision for co-orchestration systems to combine instructor-AI and instructor-learner co-orchestration (Holstein & Olsen, 2021). Features that may help address this challenge include a student-informed list of preferred and blocked partners, as well as allowing students to request a different partner or to request that the collaboration be halted when it is of poor quality. Additionally, when deciding the proper partner, timing, and content for a collaborating pair, one potential way to incorporate both teachers' and students' agency and control can be to have one party propose ideas and the other party make the final decision (e.g., teachers propose collaborating content and students choose, or students propose several partners and teachers confirm). Another general approach is to co-design with teachers and students on design ideas that might balance their preferred level of control in the classroom.

5.10 Ethical Implications

With the adoption of tools that support dynamic transitions can come ethical implications, as revealed in the current study. *Firstly*, dynamic transitions, which inherently require students to switch between different learning activities, may pose additional challenges for students with learning disabilities, as revealed in this study. Since these students may have impaired attention spans (Bateman, 1992; Vaughn & Linan-Thompson, 2003) and may have difficulty

focusing on and staying engaged in one learning activity, switching between different ones may compromise their learning effectiveness. *Secondly*, as compared to traditional group formation, which usually involves the whole class so that every student has a chance to have a partner, the nature of dynamic transitions encourages instructors only to pair students for whom *they deem* collaboration to be beneficial. While teachers express that they often leave students who *do not need help* to work by themselves, it is possible that their judgments are not always accurate and/or up-to-date. This is evidenced in my analysis, which found that students left in individual mode have lower performance than other groups. Given that dynamic transitions, as supported in this current technology ecosystem, rely quite heavily on *teachers'* subjective judgment, dynamic transitions may risk keeping struggling students in individual learning states. Thus, given that peer tutoring holds the potential to scaffold students, keeping them in individual learning may potentially prolong or exacerbate students' struggles, without careful design, such as scaffolding to help teachers spot struggling students or allow students more control in the process (Echeverria et al., 2020; Holstein & Olsen, 2021; Sharples, 2013). *Thirdly*, given that teachers may be inclined to pair students who they know would work well with their peers, one potential risk of dynamic transitions is not providing a fair and equal opportunity for collaboration, for students who teachers may consider to "*have social behavior issues*" or "*do not work well with others*". Future tools of dynamic transitions should consider these ethical implications in the tool design process.

5.11 Conclusion and Role in Thesis

I conducted a classroom prototyping study of a human-AI co-orchestration technology ecosystem. The prototyping study shows that dynamic transitions between different activities have pedagogical value in the eyes of teachers and, to a lesser degree, of students. Teachers view dynamic transitions as a way to achieve more differentiated learning that can be better customized to each student's individual pace and preferred learning mode. Supporting dynamic transitions in actual classrooms is worthwhile from a pedagogical perspective, yet not without challenges such as finding the optimal timing, content, and partner during such transitions. Especially, the challenge may lie in 1) supporting teachers in identifying the optimal moments and suitable partners, without overloading them in classroom management, 2) ensuring students are responsible, responsive, and contribute in a constructive way during collaboration, 3) allowing smooth, non-disruptive transitions that switch students to learning activities in moments they needed the most to get them unstuck, without dragging collaboration to turn into unproductive social or idle time. The current prototyping study builds a case that with Pair-Up and the two AI-based tutoring systems that support individual and collaborative learning, respectively, dynamic combinations of individual and collaborative learning are feasible in classrooms, and viewed as desirable by teachers and (to a lesser extent) students. The study also highlights a number of ways in which the design of the technology ecosystem could be improved.

This is the first prototyping study in authentic classrooms on dynamic transitions as supported by a human-AI co-orchestration technology ecosystem. The findings, insights and design implications in this study are not limited to the specific design case of Pair-Up or the co-orchestration technology ecosystem. One goal is to spur more work at the intersection of HCI and education for adopting this lens of human-AI co-orchestration. I believe there are various possibilities to embody the insights from the classroom study and the design implications in future tools that support classroom learning and teaching, or leverage complementary strengths of human and AI systems synergistically.

This study also marks an important milestone in my thesis, as this is the first classroom study that I conducted in my thesis with authentic users of teachers and students in their classes around the line of work of dynamic differentiation. This study is not only a very important realization and demonstration of whether this concept of dynamic differentiated learning can work well in actual classrooms, as supported by the technology ecosystem, but also paves the way for the next important classroom study in my thesis (Chapter 6), where I conducted a controlled experiment in classrooms, comparing dynamic versus standard transitions.

CHAPTER 6. A CONTROLLED EXPERIMENT COMPARING DYNAMIC AND STANDARD TRANSITIONS

While there are many potential merits of dynamically combining individual and collaborative learning, it is unclear how doing so might compare with *standard* ways of combining the two learning modes. Dynamic combinations of individual and collaborative learning may bring about self-paced learning experiences. Students who finish an activity earlier than their peers would not need to be idle while waiting for their classmates to finish before they can transition to another learning mode (J. K. Olsen et al., 2020). On the other hand, standard ways of combining individual and collaborative learning may also have merit. For example, the pre-planned nature may give teachers adequate preparation and lessen the in-the-moment management workload.

This study aims to bridge this literature gap by conducting a controlled experiment to evaluate and compare dynamic and standard combinations of individual and collaborative learning in 13 classrooms, where students used intelligent tutoring systems. This study leverages a smart technology ecosystem that supports students with student-driven, self-paced problem-solving and teachers with data-driven, real-time learning analytics. Through mixed-method analysis of post-study survey, interviews, tests and system log data, this study contributes to HCI research by investigating 1) students' learning outcomes and processes in conditions where they switch between individual and collaborative learning in a dynamic or standard way; 2) the respective classroom stakeholders' perceived value of these different combinations of individual and collaborative learning supported by the co-orchestration technology ecosystem, and 3) design opportunities and challenges for future tools.

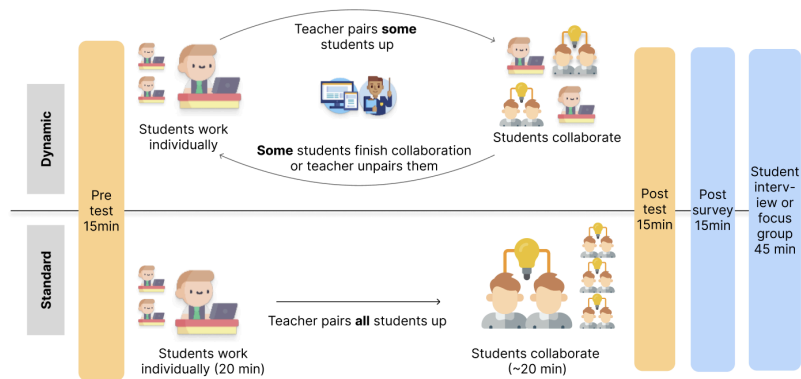


Figure. 6.1: Study Experiment Conditions and Procedure: Upper: Dynamic Condition; Lower: Standard Condition

6.1 Research Questions

Through an experimental study, I evaluate the differences between the dynamic combination versus the business-as-usual (standard combination) of collaborative and individual learning. The study aims to answer the following research questions:

RQ1. Is there any difference between the Dynamic condition and the standard condition regarding students' learning gains?

RQ2. Is there any difference in dynamic and standard conditions during the students' learning process?

RQ3. How do teachers perceive a dynamic and standard combination of individual and collaborative learning?

RQ4. What are students' opinions of a dynamic and standard combination of individual and collaborative learning?

Given this is unpublished work, the thesis document contains only the high-level takeaway and omits the detailed procedure, methods, and analysis.

6.2 High-level Results

The design of the co-orchestration tool is aligned with the one prominent goal of “guiding instructors’ attention to learners who may be most in need of human attention during ongoing learning activities”. Work by Olsen et al (J. K. Olsen et al., 2020) found that carefully designed co-orchestration support for transitions was critical, so that instructors would not be overwhelmed as they simultaneously focused on providing direct instructional support. Such support allows for “more elasticity in the timing between activities, which has shown beneficial when changing social dynamics” (Manathunga et al., 2015).

Despite the proliferation of educational technologies aimed at optimizing teaching and enhancing student learning (e.g., (Alavi & Dillenbourg, 2012; Cuendet et al., 2011; Holstein, McLaren, et al., 2018; Mercier, 2016; P. Wang et al., 2018)), many primarily focus on either individual or collaborative learning, often overlooking the fluid transition between these modes (J. K. Olsen et al., 2020). To address the difficulties teachers encounter in seamlessly integrating both learning approaches without digital support, I collaborated with educators to develop a teacher-oriented orchestration tool (anonymized for peer review). This study leverages this smart technology ecosystem that supports students with student-driven, self-paced problem-solving and teachers with data-driven, real-time learning analytics.

Recent literature has found that dynamic combinations of individual and collaborative learning can be feasible in classrooms, but there has not been a study that shows how they

compare to a traditional, business-as-usual, pre-planned way of learning that combines individual and collaborative learning. It was unknown before the current study how students' learning gains and processes might differ between dynamic and traditional combinations of individual and collaborative learning.

Practical Implications: The findings from this research suggest pathways for integrating individual and collaborative learning within future smart classrooms. They highlight the viability of such integrative approaches, while also emphasizing the need for further experimentation to refine the tools used by teachers and students. By triangulating both quantitative log data that captures the student learning process and outcomes, and user-centered insights from classroom stakeholders, this study offers valuable experimental outcomes and design perspectives for future researchers and tech professionals.

Theoretical Contributions: This study delves into the inquiry of whether short intervals of collaborative learning are most effective when dynamically allocated. Specifically, it questions if collaborative learning serves as an optimal real-time remedy to noticeable student struggles, particularly within the context of intelligent tutoring software. Although the research doesn't conclusively address this question, it hints at the potential advantages of such an approach. The study also validates an instructional practice noted in the prior classroom observations using a distinct analytics-driven teaching tool: the spontaneous pairing of students by teachers when a significant number face challenges (anonymized). Previous research studied dynamic shifts between various social levels (J. K. Olsen et al., 2020). The most comparable prior work is a prototyping study that established the feasibility of combining individual and group learning within a tech-driven ecosystem (anonymized). To the best of my understanding, this present study stands unique in empirically assessing student learning outcomes and processes across both dynamic and conventional, pre-planned methods of combining individual and collaborative learning.

6.2.1 Learning Gains & Process in Dynamic and Standard Conditions

Learning Gains (RQ1) From the analysis of pre- and post-test scores, students demonstrated an enhancement in procedural test items. No notable improvement was seen in the conceptual test items. I found no difference in pre-post test gain between conditions. Independent of the conditions, role-based differences in learning gains emerged. Students in the role of solvers, only in collaboration, exhibited a greater extent of learning gains compared to those in the Tutor role for conceptual item 2. My takeaway from this analysis suggests, for at least some of the items, 1) Standard condition had more learning gains than Dynamic condition, 2) engaging students in problem-solving directly may potentially foster greater conceptual knowledge compared to when they play a tutoring role.

Learning Process (RQ2) In individual learning mode, the log data indicated no notable differences between the two conditions regarding students' efficiency in equation-solving or their correctness rate per step. However, a different pattern emerged in collaborative settings.

Both conditions saw students dedicate similar time durations to collaboration, but a distinct difference exists in the number of problems they addressed. The Dynamic condition's students tackled fewer problems than those in the Standard condition. Consequently, the average time students in the Dynamic condition allocated to each problem exceeded that of the Standard condition, indicating a possibly higher collaboration efficiency within the Standard condition. An alternative explanation can be that the students in the Standard condition started with easy problems and progressed to hard ones in collaboration, compared to the Dynamic condition, which received teacher-assigned problems, presumably more customized to the learning level of the given student pair.

Furthermore, in terms of accuracy, the Standard condition exhibited a higher average correctness rate compared to the Dynamic condition. This aligns with a previous study by Olsen et al., which captured a sentiment among teachers: *"It is a ton of frontloading to get them to collaborate, but once it is up and running, it is truly beautiful"* (J. K. Olsen et al., 2020). A single switch was made in the Standard condition, allowing all students to collaborate on an extended assignment. In contrast, the Dynamic condition involved more frequent switches with shorter collaborative tasks, averaging around 1.8 collaborative student assignments.

While the Dynamic condition's constantly changing tasks might appear more engaging to some educators, my findings revealed that students encountered significantly more confusion in the Dynamic condition when transitioning between individual and group learning. The advantage of the Standard condition might stem from fewer task-switching instances and reduced preparation time for new activities. This minimized task-switching potentially aids students in honing their focus, permitting consistent practice on similar questions, which may increase efficiency and accuracy.

6.2.2 Teachers and Students Perceived Value in Combining Collaborative and Individual Learning

Conceptual boundaries of orchestration research often center around "real-time management", and the role of the instructor (Holstein et al., 2019b). Through the lens of orchestration, the "usability and effectiveness of learning technologies are considered at the level of the whole learning environment, and the focus is not on any one learner, but rather on how a network of learners, instructors, and technologies interact" (Holstein et al., 2019b).

In my study, teachers and students demonstrated a subjective preference for the Dynamic condition, which stemmed from the flexibility and individualization that the Dynamic condition offers. Teachers appreciated the increased autonomy it provided, allowing them to decide if, when, and how to pair students, as well as whom to pair. This autonomy easily accommodated circumstances such as student absences and facilitated reciprocal peer tutoring. Teachers noted that since the collaborative assignment is short, students are less picky and more open-minded about exploring different task partners. Students may also be

less distracted by the need to figure out who their partner is (from the other side of the computer) as compared to in a longer assignment in the Standard condition. One potential concern teachers mentioned is that the Dynamic condition might require multitasking and potentially more classroom management load. As for ideas for personalized learning, I found teachers like customized problem sets for both individual and collaborative learning, and are more willing to have students be more self-paced in individual learning compared to collaborative learning.

As for students, in individual learning, students favored customized problems, but in collaborative settings, they preferred working on identical problems as their classmates. Students in the Dynamic condition showed a higher preference for personalized content. Students prefer doing the same problems as their peers to enable mutual assistance, avoid confusion, and gauge their mastery levels. However, many favor customized problems due to the diverse pace of learners. Less confident students feel custom problems reduce stress, while high-achievers desire challenges tailored to their level. Customized problems also minimize cheating and similarly allow students to discern their mastery relative to their peers. Students generally like combining individual and collaborative learning because they like being able to practice by themselves while also having the option to work with partners. On the downside, students may have concerns that if their partner cannot understand them, it might hold them back, or feel lonely. Connecting to the theoretical discussion (Cress & Kimmerle, 2018), collaboration in this context occurs through direct dialogical communication (chat interface), but also through common development of shared artifacts or practice (i.e., correctly solving equations), which constitutes a continually evolving process.

6.2.3 Design Tensions and Opportunities

In this research, while classroom stakeholders (both teachers and students) showed a preference for the dynamic integration of individual and collaborative learning, it didn't necessarily translate to superior learning outcomes. Drawing from this mixed-method analysis, we've identified certain design tensions inherent in implementing dynamic combinations and personalized learning. By addressing these tensions, I propose potential strategies to enhance the effectiveness of such dynamic and personalized educational approaches.

Harmonizing Flexibility, Classroom Control, and Management Load. While teachers exhibited a clear preference for the Dynamic condition due to the increased flexibility and control it provided, there's an inherent tension between these benefits and the potential for increased management demands. Enhancing teacher control, for instance, through real-time insights into student activities, could inadvertently elevate their management burden. The challenge, therefore, lies in designing classroom orchestration tools that amplify teachers' autonomy without overwhelming them. Implementing technology that supports teacher mobility, such as tablets or mixed-reality glasses (cf. Lumilo (Holstein, Hong, et al., 2018))

could offer a solution. Additionally, seamlessly incorporating learning analytics into classroom settings might allow educators to access and leverage insights without feeling overwhelmed. For example, tools like ClassBeacons (An et al., 2018) and Lantern (Alavi & Dillenbourg, 2012) which emphasize ambient awareness, can guide the design of education technology that seeks to amplify teachers' classroom awareness and control in a non-intrusive manner.

Balancing Personalization Versus Coherent Learning Experiences with Peer Support. Students perceive personalization as a means to enhance efficiency and effectiveness in their learning. This approach enables learners to progress at individual paces, preventing stagnation on mastered topics or being overwhelmed by advanced material. On the flip side, the advantage of a synchronized classroom pace is the opportunity for immediate peer assistance during challenging moments. Recognizing that peers sometimes provide clearer explanations than instructors, future ed-tech solutions should merge the benefits of personalization with functionalities that enable spontaneous peer interactions and support, such as integrated student chat features.

It's also worth noting that when adopting personalized learning experiences, student activities and paths can diversify and become intricate. My research underscores that in such personalized platforms, especially those with frequent task-switching, students might encounter confusion. Technology designers must be careful not to disrupt students' learning progress. One viable design solution is to allow students some autonomy in deciding how such transitions happen so they are more mentally prepared, ensuring a more seamless learning experience.

Fostering Collaboration while Ensuring Effective Communication. Students appreciate the blend of individual and collaborative learning; however, they have reservations about the quality of collaboration, specifically regarding the proficiency and responsiveness of their peers. My findings suggest that students benefit from sample dialogue to guide them on how to communicate as a peer tutor, which may lead to enhanced peer communication. Specifically, students note that more frequent exposure to such guidance in shorter assignments in the Dynamic condition can be more helpful, as opposed to the singular, initial guidance provided in the standard collaborative setting.

6.2.4 Role of this study and how it connects to the thesis theme

This study serves as the final study along the line of “*differentiation*.” Following the close previous study, which demonstrates the feasibility of achieving dynamic differentiation and a combination of collaborative and individual learning as supported by the technology ecosystem (chapter 5), this study (chapter 6) takes it one step further, seeking to prove through empirical analysis of the value of dynamic transitions, as compared to the business-as-usual way of transitioning between individual and collaborative learning.

While there are many possible forms of differentiated learning, one reason that I study this particular form of differentiated learning (i.e., dynamic combination of individual and collaborative learning) is that initial prior research has already found promising results. Prior work showed that a combination of individual and collaborative learning can be more effective than either mode alone (J. K. Olsen et al., 2019). However, despite the promising potential of such a combination, previous research has not achieved an actual technological realizable that allows the teachers and students to engage in such differentiated learning in an autonomous and independent way without researchers' facilitation (c.f. previous work has investigated such dynamic combinations of in classrooms in a wizard-of-oz way that require researchers to play the role of the AI system) (Echeverria et al., 2020). Practically, we contributed to a technology ecosystem where such a dynamic combination can be feasible in the classroom and where future research can build on or draw insights from. Theoretically, we contributed teachers' preferences, boundaries in human-AI co-orchestration, insights and opinions from teachers and students of such a dynamic combination after they experienced it first-hand, as well as empirical analysis results from the learning process and orchestration process, that deepen our understanding of this form of dynamic differentiated learning.

III. REFLECTION

Acknowledgement. The work in Chapters 7 - 10 is supported by RETTL funding. The project members mainly include a PhD student (myself), collaborators (Vanessa Echeverria (from Monash University), Conrad Borchers (CMU), Ann-Christin (University of Ruhr, Bochum), PIs (Vincent Aleven and Nikol Rummel). This work was built on work by a post-doctoral researcher, Ahmed Butt, who left the position on May 2024. The work is also built on the foundation of work by Prof. Shamyia Karumbaiah who contributed to the work as a post-doctoral researcher, who had left the position prior to me officially joining the project. There are undergraduate or master's students who get involved in the format of either independent study or summer REU students under the supervision of the research team (including Tianze Shou, Megna Kokkalera, Lea De Vylder, Ziqi Ding, Wenhan Li, Yi Zhao, Sijia Fan, Menghan Liu). I include further acknowledgements for more details about the team's contribution at the beginning of each following chapter.

CHAPTER 7 TEACHER PREFERENCE IN BUILDING AN MMLA REFLECTION TOOL

Acknowledgement. The survey design and data collection were led by previous team members, including Vanessa Echeverria, Ann-Christin Falhs, and Shamyia Karumbaiah. The project was in collaboration with Conrad Borchers (CMU) and Ann-Christin (University of Ruhr, Bochum) guided by PIs Professors Vincent Aleven and Nikol Rummel.

While making learning environments customized to the given learning contexts and differentiated can be conducive to successful and effective learning environments, it is often needed and very crucial to reflect upon the education experience, typically after the instruction is done or periodically, so that instructors can debrief, iterate, and bring changes to their instruction to continually improve their practice. This part of the work represents the “*reflection*” theme in my thesis, with the goal of creating an educational environment where teachers can engage in data-driven reflection about their teaching practices.

Motivation of the RETTL Project. Past research has explored various ways of using learning analytics in teacher-facing tools, such as classroom orchestration, real-time assistance, lesson planning, and early warning systems (Papamitsiou & Economides, 2014; Rodríguez-Triana et al., 2017; Schwendimann et al., 2017). These tools primarily benefit teachers by providing student analytics (Schwendimann et al., 2017). However, few projects have focused on using analytics to support teacher reflection (Roschelle et al., 2020; Sergis & Sampson, 2017). A report concluded that systems that help teachers learn by reflection are still rare and often not tailored to teachers’ needs: “*These [systems] often fall short of being usable, friendly, or instrumental for teacher’s work. They fall short of the idea of augmenting the teacher’s intelligence and helping the teacher to grow, and often only make more work for teachers*” (Roschelle et al., 2020).

Analytics from Intelligent Tutoring Systems (ITS) have been effective for instructional purposes like discussing online homework results in class (Heffernan & Heffernan, 2014; Roschelle et al., 2016), lesson planning (Xhakaj et al., 2017), and real-time awareness of student progress and struggles (Holstein et al., 2019). Yet, it remains unexplored whether combining ITS analytics with other modalities like video, location data, and physiological measures can support collaborative teacher reflection on classroom practices. Multimodal analytics could provide valuable evidence of students’ learning and teachers’ interactions, enhancing teachers’ awareness of their practices.

It is challenging for teachers to notice and remember all classroom activities (Van Es & Sherin, 2002). Learning analytics from ITS can present evidence of student challenges and struggles, as seen in teacher dashboards (Holstein et al., 2018a; Xhakaj et al., 2017). Multimodal analytics could reveal trends in teacher monitoring strategies that are otherwise

unnoticed (e.g., movement patterns, time spent with students; Holstein et al., 2017b). They may also reveal hidden relations, such as how student behaviors depend on teacher movement or stress levels. Analytics-guided video snippets can highlight interesting episodes worth reflecting on. Generally, analytics support an evidence-based approach to reflection, helping teachers evaluate practices taught in professional development (PD) (Marcos et al., 2011). As a way to engage in lifelong professional learning, K-12 teachers may reflect on their own and others' teaching practices (e.g., Clarke & Hollingsworth, 2002; Dewey, 1904; Korthagen, Loughran & Russell, 2006), including students' challenges and how to address them.

Based on this, we design an analytics tool that scaffolds collaborative teacher reflection, based on theoretical models, e.g, Theory of Change Model (Fig 7.1, Fig 7.2).

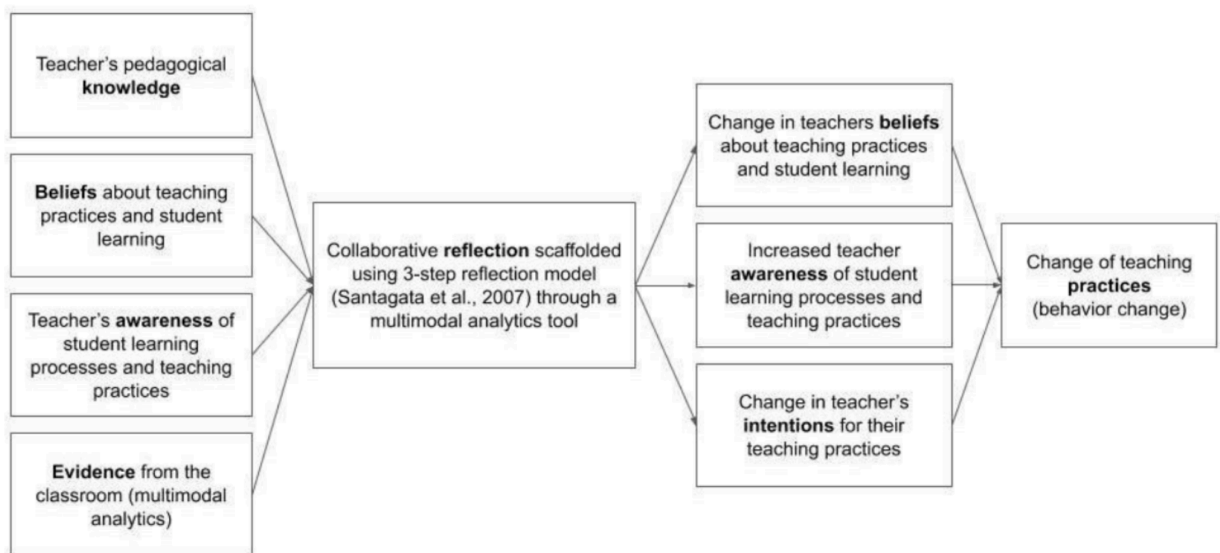


Figure 7.1. Theory of Change Model (From the RETTL Funding Proposal)

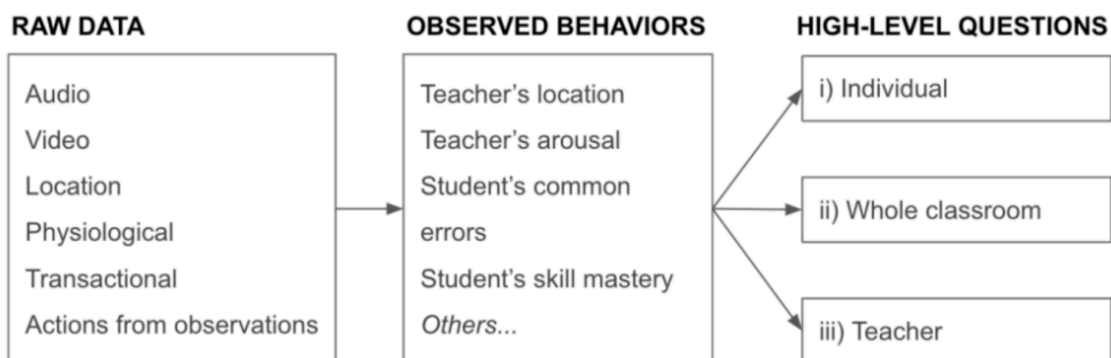


Figure 7.2. Multimodal Data, Observed Behaviors, and High-Level Questions for Teachers' Reflection

7.1 Background and Motivation

In preparation for the design of such a multimodal teacher reflection tool, I conducted data analysis on a teacher survey to understand teachers' preferences for using MMLA for reflection.

MMLA promises to improve educational technology design that supports teacher reflection. MMLA seeks to provide a more holistic view of the classroom, as it may capture traces of teacher and student behaviors (such as student engagement, performance, cognition, and emotional states), thereby enabling a more informed approach to teachers' reflection on instruction and learning. While analytics can offer valuable insights into teaching and learning processes, there is a significant gap in understanding which data elements and analytics are most beneficial for teachers (Borchers et al., 2023; Yan et al., 2022).

Recent literature reviews reported that MMLA solutions often focus on developing computational models, capabilities, and analytics for research purposes rather than on practical applications that close the analytic feedback loop (Alwahaby et al., 2022; Yan et al., 2022). Additionally, implementing MMLA tools often encounters challenges, including stakeholder preferences, privacy concerns, and the reluctance to share sensitive data (Alwahaby et al., 2022; Yan et al., 2022). For example, Cukurova et al. highlight that MMLA, venturing into complex territories with scarce guidelines and regulations, needs special consideration for vulnerable populations, such as children, so as not to contribute to a surveillance culture or raise privacy concerns (Cukurova et al., 2020; Ouhaichi et al., 2023). Still, given the potentially invasive nature of using multimodal data and AI in classrooms, there is a lack of in-depth discussions of ethical and practical considerations when designing and using MMLA. Involving teachers in the design of intelligent tools can facilitate MMLA technology adoption in the classroom. Studying teachers' considerations of MMLA analytics and data collection practices in classrooms allows us to compare the acceptance of different data modalities needed to generate MMLA. The acceptance of data collection methods and modalities might not always overlap with what downstream analytics teachers are interested in.

My research probes these challenges in the context of designing a teacher reflection tool. I address three research questions: RQ1: What analytics on *teaching* and *student learning* do teachers want to use in reflection practices, and why? RQ2: What data collection modalities on *teaching* and *student learning* would teachers allow in their classroom, and why? RQ3: With whom are teachers willing to share analytics about *themselves* and *their students*, and what privacy concerns might they have, and why?

Advancing teacher reflection tool design, I aim to identify teachers' preferences and effectively integrate analytics into educational practices in a manner that respects privacy and fosters an environment conducive to learning. This study seeks to pave the way for the development of analytics for teacher reflection, to enhance teaching, and learning and to discuss the ethical considerations inherent in using MMLA data.

7.2 Methods

Participating teachers were recruited using Prolific. Teachers were asked to consent before participating in the study and received \$20 USD as compensation. The sample included 100 teachers ages 20-62 ($M = 38.15$, $SD = 11.11$), including 64 female teachers, 35 male teachers, and one non-binary teacher. Teachers were located in the US, had English as their primary language, and had taught in primary, middle, or high school. Additional teacher demographic information is in Table 7.1.

Table 7.1. Teacher Professional and Demographic Data Based on Survey Responses

Data Element	Teachers' Information
Degree Obtained	4-year professional degree (90), PhD degree (7), 2-year college degree (3)
Years of Experience	More than 8 years (57), 7-8 years (6), 5-6 years(12), 3-4 years (13), less than 3 years (12)
School level taught	Primary school (12), middle school (45), high school (41), specialized tutoring (2)
Subjects	Language (47), math (34), science (30), and other subjects, including art, music, and history (21)
Experience with Edtech software	Half of the teachers (48) had used an educational software (e.g., ALEKS, Khan Academy) to support student learning

7.2.1 Survey Questions Corresponding to Research Questions

The survey is based on literature at the intersection of teacher development, reflection practices, and multimodal data in classrooms (Cukurova et al., 2019; Marcos et al., 2011; Ouhaichi et al., 2023; Prieto et al., 2020). Before deployment, two PhD students conducted pilot sessions and refined the survey. The survey was hosted online using Qualtrics. The survey lasted approximately 40 minutes. Teachers were prompted to imagine that they have an intelligent augmenting tool capable of observing their classroom and then using that information about their class to support their reflection. In that way, the survey provided accessible language to explain MMLA to participants.

The first research question (RQ1) related to teachers' preferences for specific analytics about themselves and their students in the context of reflection. Teachers were asked in the survey:

“What kind of information about teachers (including yourself) would you like to have access to when reflecting? Imagine you have superpowers and can easily observe and pull out information about your classroom”. Teachers selected their top three preferences out of nine options and then ranked them (Table 7.2). Similarly, teachers ranked their preferences for student learning constructs out of nine options concerning the usefulness of constructs for teacher reflection (Table 7.3).

The second research question (RQ2) related to teachers’ acceptance of data collection modalities related to teaching and student learning. Teachers ranked four modalities on their level of acceptance: audio, video, location data, and log data from interactions with learning systems (Table 7.4). Teachers then separately rate both data collection modalities for teacher data and student data (there is an additional option of a smartwatch in gathering physiological data for teachers’ data, which was excluded from the analysis to ensure comparability of rank distributions between teacher and student data).

The third research question (RQ3) related to teachers’ willingness and concerns regarding data sharing about themselves and their students. Teachers responded to seven checkboxes for different stakeholders (Table 7.5), representing their willingness to share data and analytics of interest regarding their students’ learning and teaching practices. It was possible to check all or none of the options. The data collection modalities related to data sharing were limited to what teachers had previously indicated they would be willing to collect, assuming that teachers would not be willing to share data they do not find acceptable to collect in the first place.

7.2.2 Analytical Methods

Quantitative Survey Data. For RQ1, I report how often teachers chose a given analytic item among their top three preferences, including 95% binomial confidence intervals, to estimate uncertainty regarding preferences in the sampled population. I chose this statistical measure over median rankings due to considerations of statistical power: while probabilities of being chosen by teachers in their top three are less fine-grained than estimations of specific ranks, they can be estimated with higher statistical certainty.

For RQ2, I compared the rank distributions of teacher preferences for student data and their own. I report teachers’ median acceptance rank for each of the four data items to be ranked separately for data about students versus data pertaining to teachers. Medians are more appropriate for ranked data than means, as ranks cannot be assumed to be normally distributed. I also reported whether teacher preference rankings were significantly different for student data compared to teachers’ data, based on two-sided nonparametric Wilcoxon signed-rank tests.

For RQ3, I calculated estimates of how likely teachers were to select each stakeholder for data sharing. For each comparison, I also conducted a nonparametric, two-sided binomial test (well-suited as it does not assume distributional properties). The goal was to ascertain whether teachers' willingness to share data for the same stakeholder significantly differed for students compared to teacher data.

Qualitative Survey Data. I collected open-ended text responses from teachers on questions where teachers explained why they selected their three most preferred analytics options and their rankings. Three coders employed the thematic analysis approach to analyze open-ended responses. Thematic analysis involves identifying, examining, and interpreting patterns or themes within qualitative data. I reviewed the responses to distill and categorize key themes and insights. I coded and analyzed each response individually through affinity diagramming using an online board (i.e., Miro). Three coders cross-checked each other's coding to enhance the reliability and reduce individual bias.

7.3 Results

7.3.1 RQ1: *What analytics on teaching and student learning do teachers want to use in reflection practices, and why?*

Preferences for teacher analytics. Table 7.2 lists the frequency of teachers, including each teacher's analytics in their top three analytics for reflection. Teachers were most commonly interested in data about ways of motivating their students (44%) and interacting with students (43%). *"Teachers have to be able to keep their students engaged by keeping them motivated. Without motivation most of what the teacher explains will be lost"* (T78). Some teachers believed student motivation is linked with teachers' emotions and classroom environment, and want to know how to effectively motivate their students: *"A teacher's emotions are sensed by the students, even when you think they do not. The mood of the teacher will set the tone for the class, so it is important to be positive and supportive [...]"* (T34). As to how to motivate students, some teachers found it important to understand students' emotions and socio-emotional learning: *"To motivate and help students, it's important to get down to their level and join their world by seeking to understand where they come from first [...] I like to prioritize social-emotional development and understand where a student is coming from (because as people, I essentially make choices based upon our feelings)"* (T36). Teachers were also interested in analytics about their responses to students asking for help or feedback (35%), and general instructions to students (33%). Teacher T4 pointed out that clear, well-structured instructions are crucial in preventing misunderstandings and maintaining a positive learning environment. Teachers were particularly interested in seeing and reflecting on their responses to student requests for help or feedback. Teachers expressed interest in being aware of noteworthy events in the classroom that might have escaped their notice, for example, incidents such as cheating or student mistreatment, which could occur without their

knowledge: “I would love to know if someone slyly pulled out a vape, used it, and put it away without me noticing. I would love to know if someone got away with cheating, and I would also want to know if someone mistreated someone else without me knowing about it” (T58). Teachers were least interested in their questions to students (18%) or their emotions during teaching, such as stress and frustration (15%). Many teachers viewed collecting data about their emotions as unnecessary, privacy-invading, and limiting: “This data also feels like it could be used against me and might make me more self-conscious and attempting to limit the ways in which I express myself” (T38).

Table 5.2. Teacher Preferences for Teacher Analytics Expressed as Probabilities of Inclusion in Top Three Choices, With 95% Confidence Intervals

Data Item	Estimated Probability	CI 95%
My ways of motivating students	44%	[35%, 54%]
My interactions with students	43%	[34%, 53%]
My responses when students ask for help or feedback	35%	[27%, 45%]
My instructions to students	33%	[24%, 42%]
Noteworthy events that I was not aware of	31%	[23%, 40%]
How I monitor and observe individual students	22%	[15%, 31%]
How I monitor and observe the whole class	22%	[15%, 31%]
My questions to the students or the group	18%	[12%, 26%]
My emotions (e.g., stress, frustration)	15%	[9%, 22%]

Preferences for student analytics. The student analytics that teachers were most interested in for reflection was student learning and progress (55%) (Table 7.3). One teacher explained how it connected to reflection, “*Knowing student learning and progress would help me know what the weakest and strongest areas of my teaching are. This could help me help students better, revise my teaching, and reach them better*” (T58). Teachers also favored students’ feedback on their teaching and information about students’ engagement (35%) over information regarding misconceptions, common errors, and emotions (ranging from 22% to 27%). One teacher mentioned, “*There is no way I can help a student if they are off task or not focused on the learning outcome*” (T31), indicating the necessity of understanding student disengagement for effective teaching. Teachers needed to comprehend students’ challenges and their emotional well-being. For example, “*The number one data point that is not always known is any issues the student is having. If they are in a gang, doing drugs, suicidal, had a*

death in the family, or have friend issues all play into their engagement at school” (T16). 18% of teachers picked student mastery of skills, knowledge, and progress as assessed in learning software.

Table 7.3. Teacher Preferences for Student Analytics Expressed as Probabilities of Inclusion in Top Three Choices, With 95% Confidence Intervals

Data Item	Estimated Probability	CI 95%
Students’ learning and progress	55%	[46%, 64%]
Students’ feedback	36%	[28%, 46%]
Students’ disengagement	35%	[27%, 45%]
Improved learning after I help a student	30%	[22%, 39%]
Misconceptions in learning	27%	[19%, 37%]
Students’ common errors	26%	[18%, 35%]
Students’ emotions	22%	[15%, 31%]
Information from learning software about students’ mastery of new skills and knowledge	17%	[11%, 25%]
Information from learning software on students’ progress	13%	[7%, 20%]

7.3.2 RQ2: What data collection modalities on teaching and student learning would teachers allow in their classroom, and why?

Table 7.4 describes the median ranks of teachers’ acceptance of different data collection modalities, of their own and students’ data. **First, log data emerged as the most accepted modality for student data.** Teachers think students’ log data helps them reflect, as, *“behaviors while using the educational software would help me understand if I am monitoring students enough or in the right way”* (T46). Some teachers, though, had concerns about not knowing how to effectively leverage and use student log data. Student log data was followed by a tie of audio and video data of students. Student location data was the least accepted among all options. In contrast, teachers’ general preferences for different data collection modalities for *teacher data* were more mixed, lacking a clear trend or consensus among teachers as to what type of data collection they would accept or feel comfortable with. Specifically, all four data collection modalities ranged from median ranks of acceptance between 2 and 3, which are around the middle of the 1-4 range.

Table 7.4. Teachers' Acceptance of Data Collection Modalities as Median Ranks, Separately for Student and Teacher Data, With Two-Sided Pairwise Wilcoxon Rank Tests

Data Type	Rank (Student Data)	Rank (Teacher Data)	<i>p</i>
Log Data	1	2.5	.210
Audio	3	2	.002
Video	3	2.5	.109
Location Data	4	3	<.001

We conducted two-sided pairwise Wilcoxon rank tests to determine if teachers were significantly more accepting of each data collection modality for their students' data than their own. Two significant differences emerged. **First, teachers were significantly less accepting of collecting student location data compared to their own location data ($p < .001$).** From qualitative analysis, teachers had quite polarizing opinions regarding collecting location data. Some teachers thought it might **help teachers reflect on their teaching behaviors**, or found it intriguing, e.g., *"Knowing how I moved around the classroom could be very helpful regarding classroom management, such as places I'm missing consistently or places that I visit too frequently. Do I move around too much or too little, etc?"* (T28). On the other hand, many teachers had **privacy concerns over using teacher location** and strongly disagreed about collecting location data in school: *"I hate the idea of something similar to GPS because I feel like that is way too much monitoring with little control"* (T93). Other reasons for reluctance included teachers **not seeing its relevance to learning**: *"I think information related to position or software interaction is the most limited because they share the weakest correlation with teacher efficacy in my opinion"* (T44). Some teachers expressed that it may not be helpful for a small class size, as teachers may already know their routes and visit frequency. The survey kept it open on how the location tracking happens, thus teachers may have different assumptions (e.g., many assume GPS collection).

Second, **teachers were significantly less accepting of collecting audio data of their students compared to audio recordings of themselves ($p = .002$),** whereas such a difference for *video* was not found. The primary advantage teachers saw in video data was that it contained the most comprehensive information, *"...including emotions, verbal instructions, and other features of teacher-student interactions"* (T44). Teachers also thought videos could remind them of things they may have missed during class. *"Things are easily missed when attention is split across an entire classroom, and this can point out some of those missed moments"* (T68). The disadvantages of video data that teachers saw related to privacy concerns, and teachers may be uncomfortable with video recording in their class. As for audio, the advantage was that teachers thought it could remind them what happened in class:

“I think audio is going to be helpful because then you can listen back to the lesson and the questions asked and reflect on what went well and what they need more help on” (T22). Teachers thought audio data may have fewer privacy concerns than videos, though it contains less information.

7.3.3 RQ3: *With whom are teachers willing to share teacher and student data?*

I asked what stakeholders teachers would be willing to share their own and their students’ data with (where teachers could check all or no boxes) and computed the percentage of teachers willing to share data with each particular stakeholder called out in that question. The results are summarized in Table 7.5.

First, regarding teachers’ general levels of data sharing acceptance, **principals emerged as the most accepted stakeholders with which data was shared, both for student (56%) and teacher data (41%)**. Legal restrictions also play a role in data sharing: *“Sharing students’ data with other people except the principal is not legally appropriate”* (T60) or *“We are only allowed to share student data with the administration”* (T43). **Regarding their own data, teachers were more comfortable sharing data with other teachers teaching the same subject (29%) than with any other colleague in their school (15%) or colleagues in the same teaching year (18%)**. A similar relative ranking emerged for the sharing of student data. Notably, teachers were more likely to share their data with coaching experts over subject coordinators (27% compared to 20%) while the opposite was true for student data (26% compared to 23%).

Table 7.5. Teachers’ Willingness to Share Data With Different Stakeholders, Expressed as Percentages of “Check All That Apply” Responses, With p-Values From Two-Sided Binomial Tests

Data Element	Acceptance (Student data)	Acceptance (Teacher data)	<i>p</i>
No one	14%	26%	.002
Any colleague who works at the school	18%	15%	.293
Colleagues from the same teaching year	18%	18%	.999
Coaching expert	23%	27%	.463
Subject coordinator	26%	20%	.104
Other colleagues from the same subject	38%	29%	.030
Principal	56%	41%	.002

Comparing the relative acceptance of sharing teacher data compared to student data with the same stakeholder, **teachers were generally less willing to share their data (26%) than their students’ data (14%) based on the “(share with) no one” option**. This difference was statistically significant ($p = .002$). Many teachers believed sharing students’ data needed consent (e.g., from students and parents). 38% of teachers were willing to share students’ data with colleagues from the same subject, and significantly fewer teachers were willing to share data about themselves (28%, $p = .030$). Teachers thought data sharing with colleagues might

benefit teaching, and especially preferred data to be de-identified: *“Sharing with colleagues, particularly if done without identifying information, could help me in receiving feedback for making improvements”* (T11). Similarly, **teachers were significantly more willing to share their students’ data compared to their own data with their principal ($p = .002$)**. Teachers thought sharing students’ data with the principal could help support teachers: *“I think the principal, as the head of the school, would find the information valuable for making informed decisions and supporting teachers effectively”* (T70).

7.4 Discussion and Study Takeaway

Multimodal Learning Analytics (MMLA) seeks to leverage multiple sources of data in high frequency (e.g., video, log data, audio, and position data) to provide a more holistic view of the classroom (Giannakos et al., 2022). MMLA captures traces of teacher and student behaviors and holds promises to inform educational technology design that supports teacher reflection. Despite the MMLA’s potential to generate rich analytics through classroom sensors and TEL learning environments, there is a lack of research that designs MMLA solutions around teacher preferences, needs, and concerns (Alwahaby et al., 2022; Yan et al., 2022). To address this gap, the present study investigates teachers’ perceptions and preferences regarding the potential use and data sharing of MMLA to support reflection on their teaching practices.

My first research question (RQ1) asked what analytics teachers would prefer most. **For teacher analytics, teachers expressed the strongest interest in data about their ways of motivating students and interactions with students.** This teacher preference poses a methodological challenge to accurately capture students’ motivation and attribute changes in student motivation to teacher-student interactions. Although challenging and requiring rich interaction data, prior work has proposed methods to analyze teacher discourse via TEL (e.g., (Cao et al., 2023)). Teacher-student conversations could be analyzed through similar methodologies to speak to how positive, inclusive, or conducive to students’ motivation interactions are. However, a more fundamental research challenge may arise from an incompatibility between teachers’ preference *not to* capture student audio and location data (needed for detailed descriptions of teacher-student interactions) and their desire to view analytics about teacher-student interactions. A middle ground could be to capture teacher position data only at the expense of richer descriptions of interactions, for example, by measuring how commonly teachers visit particular students. **For student analytics, teachers expressed the strongest interest in knowing students’ learning and progress,** but somewhat surprisingly expressed the least interest in seeing students’ skill and mastery of the software. While students’ skills mastery data has been used in intelligent support for learning, teachers may not relate this data to their teaching practices, potentially due to teachers having limited knowledge or training in interpreting software log data and correlating it with

students' skill mastery. As an alternative interpretation, the discrepancy could also relate to teachers in this sample not trusting the accuracy with which the tutoring software estimates learning or could be a reflection of excluding the (presumably less accurate) estimation of learning if the construct of learning itself is already included in teachers top three choices (which would mean that this finding can be attributed to the survey methodology). More qualitative investigations (e.g., via follow-up interviews) could elucidate this issue further.

For RQ2 (related to what data modality teachers accept for data collection), I found that **teachers were generally most open about log data collection and much more accepting of students' log data than teacher log data**. Teachers were least accepting of teacher and students' location data (a relatively new modality, (Alwahaby et al., 2022; Yan et al., 2022)), primarily due to privacy concerns or a lack of knowledge on how location data can be relevant to learning. Prior work, however, has shown promise (e.g., (Karumbaiah, Borchers, Shou, et al., 2023)), for teachers to distill aspects relevant to student learning by viewing their location data analytics and reflecting on their teacher visits. This difference between recent research trends and teacher preferences is notable, pointing to a potential neglect of MMLA research trends to incorporate teacher preferences into research priorities: if teacher preferences are not considered at an early stage of analytics development, eventual adoption of TEL tools supported by MMLA could suffer.

For RQ3 (teachers' data sharing preferences), **teachers are most willing to share data with principals and colleagues teaching the same subject (e.g., math)**. Sharing with other colleagues, however, is decidedly less preferred. **Teachers are also less willing to share their data than their students' data**. Combining RQ2 and RQ3, teachers find it *both desirable and acceptable* to use students' log data, which can convey student learning and progress, which interests teachers. Notably, there was no clear modality that teachers accepted most or least regarding the collection of their own data. This could point to a potential lack of understanding of nascent MMLA research practices and what each teacher's data modality would entail when collected in K-12 classrooms. I gathered some evidence for this interpretation based on teachers' open-ended responses related to teacher log data, for example, which some teachers were unclear about. If this interpretation is true, then more is to be done to convey to classroom stakeholders the potential value of these different teacher data modalities and collect MMLA in a manner that respects classroom norms.

The present study revealed surprising insights that could inform TEL research practices. **First, for student analytics, teachers were most interested in seeing student learning situations but expressed little interest in accessing information from learning software about student mastery and progress**. As discussed, this finding could open research opportunities to study further if teachers are unable to relate learning analytics to mastery and progress in their teaching practices or lack trust in such analytics (irrespective of knowing

how to interpret them). Second, I found **teachers favored seeing students' engagement and overseeing their misconceptions and common errors**, potentially due to teachers in this sample considering engagement to be more fundamental for learning or more relevant for reflection. Third, teachers were more protective of their students' data regarding some data **collection** modalities compared to data about themselves (audio and video). This finding supports the need for guidelines to preserve children's privacy (Cukurova et al., 2020). Still, teachers are more willing to **share** data about their students than their own data (specifically with colleagues from the same subject domain). From qualitative results, teachers think sharing student data can help school administrators improve education, but sharing their own data may raise concerns about self-consciousness, risking creating a surveillance culture or hurting teachers' teaching autonomy.

Implications on Reflection Tool Design. My study found that teachers found it both desirable and acceptable to collect students' **log data** to support teachers' reflection. There is, however, usually an abundance of log data, and the design challenge for teacher reflection tools is to develop 1) intuitive representations that teachers can make sense of and find interpretable, and 2) simple visualizations that teachers can quickly glance at. A human-centered perspective is needed, and with the rapid advancement in natural language processing, TEL research increasingly uses audio data to understand the learning process and create teacher-facing analytics (Pugh et al., 2022). From this study, teachers found **video and audio data** helpful and may support them to reflect on how they motivate students and whether they are getting the required attention. Processing video and audio data, however, can be time-consuming. One practical design challenge is understanding what teachers consider noteworthy events and how to capture these *key events* in video and audio, which AI models and generative AI agents could help detect. Finally, given teachers' privacy concerns, one idea could be using AI models that can mask faces in video to reduce privacy concerns (e.g., (Li et al., 2023)). As for **location data**, given its nascent stage, researchers may need to convey to classroom stakeholders its potential values in capturing teacher-student interaction for TEL and its relevance to student learning (e.g., it may show whether struggling students have been adequately attended to and visited). Researchers also need to decrease privacy concerns by safeguarding data and collecting them less intrusively (Borchers et al., 2023; Karumbaiah, Borchers, Shou, et al., 2023).

One potential limitation of my study is that the sample represents US teachers, who might have different cultures and attitudes toward data sharing and collection (e.g., compared to Dutch teachers (Jessica Vitak, Yuting Liao, Anouk Mols, Daniel Trottier, Michael Zimmer, Priya C. Kumar, Jason Pridmore, 2023)). Furthermore, teachers might have interpreted the listed data collection modalities and analytics differently depending on their preconceptions about them. More work is needed to study teachers' preconceptions about different data collection modalities, which presents an opportunity to improve how TEL communicates and

shares its nascent research with the broader public.

How this work connects to the thesis theme. This work serves as the first step in my thesis to delve into how to make learning environments more reflective through data-driven insights. Specifically, I would leverage the teacher preferences and needs for analytics for reflection revealed in this chapter and connect to Multimodal learning analytics (MMLA) in the context of classrooms where students use intelligent tutoring systems. There could be many ways AI or AI-based analytics could support teacher reflection. My approach would depend on the generation of students' log data from intelligent tutoring systems, as well as the multi-modal analytics. I chose this topic because it would be interesting for me to delve into the new field of MMLA. Teachers expressed (details in chapter 5) that they are most interested in seeing teacher analytics about how they motivate students and teacher-student interaction, where MMLA is a good venue for studying this (as audio and location) may convey how teachers' visit patterns are like, or how teachers converse with students. This work also paves the way for a later stage of the tool design as well as the classroom study.

CHAPTER 8 REFLECTO - LEVERAGING MULTIMODAL LEARNING ANALYTICS IN TEACHER REFLECTION TOOL

Acknowledgement. While I participated in the user research and simulation sessions part (Chapter 5), the design and implementation of it primarily started during the summer of 2024, while I was on an internship, by Prof. Vincent Aleven and other designers and developers in the team. It builds on the technical foundations of visit detection and the pipeline established by Conrad Borchers, Peggy Shen, Lucia Fang, and Prof. Shamyia Karumbaiah with her collaborators. The user testing, tool setup, and other tasks were helped by research volunteers (including Mandy Pan, Manqing Yu, Ashley Xu, Maria Ma, Yvonne Zhao, Yikuan Fang, Liyi Xu).

Literature Motivation. Prior research has found that teacher reflection is important for both teachers' professional development and students' learning (Marcos et al., 2011), but frequent, effective reflection faces many obstacles (e.g., it is hard to be aware of all that happens during a busy classroom session) (Yang et al., 2024). As detailed in Section 2.4, there is a growing interest in using novel sensing technologies (MMLA) to automatically analyze classroom activity traces, to model behaviors such as engagement and teacher-student interactions (Martinez-Maldonado et al., 2020). MMLA has been used to study complex teaching and learning processes, to provide a more comprehensive understanding of learning and teaching (Cukurova et al., 2020; Echeverria et al., 2024). Advancements in MMLA have the potential to enhance the development and foster reflection on collaboration dynamics in physical learning environments. For example, TeamSlides and EduSense both leverage MMLA to facilitate reflection (Ahuja et al., 2019; Echeverria et al., 2024). Intelligent tutoring systems

(ITS) have a track record for effective learning and have been widely used in the classroom (du Boulay, 2019; Kulik & Fletcher, 2016; Ma et al., 2014). They provide interaction logs, a good source for MMLA to understand students' behaviors, performance, and learning situations. There is a lack of adequate prior work in the application of an integrated use of teachers' location data and students' learning data from knowledge-traced tutoring software into the practice of teacher reflection. In this chapter, I discuss the work on the MMLA teacher-reflection tool (Reflecto). Reflecto is an MMLA-based tool that was designed and implemented, with functions that log teachers' positions and allow teachers to reflect based on the analytics displayed on it. It is designed to support teachers in reflecting (after-the-fact) on their classroom practices during sessions when students are engaged with intelligent tutoring software. Echoing the broad theme keyword of *reflective*, the goal of this work is to see how the Reflecto tool helps teachers in achieving more *reflective* practice. More specifically, to investigate how Reflecto helps their reflection, affects their teaching, and what design implications can be derived from the study.

8.1 Reflecto - a teacher-facing reflection tool based on multi-modal learning analytics

Reflecto aims to promote teacher agency through out-of-class teacher-initiated, data-driven exploration of possible trends in their classroom practice, and it combines learning analytics and teaching analytics. It is worth noting that the teacher participants did not use the tool independently. Reflecto was introduced with scaffolding tasks to help teachers gain familiarity, along with guidance from the experimenter during the reflection sessions. The guidance (detailed in Section 9.4) included prompting teachers to select a trend, conduct data analysis using the dashboard, and consider whether there were any behaviors they wished to change.

The Reflecto tool description is based on the Interactive Event demo paper ⁹.

Multi-modal Learning Analytics Infrastructure for Reflecto. When students use AI-based tutoring software for self-paced practice, teachers are very busy monitoring the class and supporting students beyond the software's capabilities. This task is challenging – might data-informed reflection help teachers improve? Reflecto, designed through user-centered design with teachers (Karumbaiah, Borchers, Falhs, et al., 2023; Yang et al., 2024), supports post-class reflection by allowing teachers to explore multimodal classroom data. How do we choose what specific type of MMLA to leverage in the reflection tool? ITSs, which have been widely and effectively used in the classroom (du Boulay, 2019; Kulik & Fletcher, 2016; Ma

⁹ Part of the content of 8.1 is based on the following Interactive Event paper: Yang, K. B., Zhao, Y., Borchers, C., Fang, Y., Kim, J., Butt, A., Popescu, O., Gupte, S., Chiu, A., Shen, P., Rummel, N., & Alevan, V. (2025). Reflecto: A teacher reflection tool leveraging multimodal learning and teaching analytics. In Proceedings of the 26th International Conference on Artificial Intelligence in Education (AIED 2025), Interactive Event.

et al., 2014), can automatically capture logs for students' interaction with the systems, which holds potential to be a good source for MMLA to understand students' behaviors, performance, and learning situations. Previous research also found that teacher positioning in the classroom and proximity to students can strongly influence students' engagement, motivation, disruptive behaviors, and self-efficacy (Martinez-Maldonado et al., 2020; O'Neill & Stephenson, 2014). Combining ITSs logs and teacher position tracking, Reflecto leverages two data sources: 1) data about the teacher's location and movement through the classroom, captured by Pozyx sensing technology every second and processed to identify student visits (Shou et al., 2023), and 2) intelligent tutoring system log data, augmented with detectors for idle behavior, struggling, and hint abuse.

The use of Reflecto requires that location sensors be present in the classroom. It also requires that students' locations are logged and stable within the coordinates of the sensing system, as the visit detection depends on it (more elaborated in section 9.3.3). We leveraged the Pozyx Inc. position sensors, to register the teacher's location within the classroom. According to Pozyx performance reports, the Real-Time Localization System (RTLS) achieves an average horizontal positioning error of approximately **9.2 cm** under line-of-sight (LOS) conditions. In non-line-of-sight (NLOS) environments where some anchors are obstructed, the horizontal error increases to around **14 cm**, while vertical positioning accuracy is typically lower due to anchors being installed at similar heights (Source: <https://www.pozyx.io/creator-system-performance>). Prior researchers have used the position sensors' data to automatically record teachers' positions in the classroom and algorithmically infer visits (with reasonable albeit imperfect accuracy) (Karumbaiah, Borchers, Shou, et al., 2023; Shou et al., 2023). The accuracy of the visit detection algorithm for particular student(s) in a K-12 math classroom working with intelligent tutoring systems had a large variability in stop detection accuracy given different parameter choices (Karumbaiah, Borchers, Shou, et al., 2023; Shou et al., 2023). The teacher visit detection algorithm achieved an F1 score of **0.18** under conditional maximization (precision = 0.21, recall = 0.17), which was about twice as high as the **0.09** F1 score under absolute maximization (precision = 0.05, recall = 0.68). This tradeoff shows that while absolute maximization favored high recall at the cost of very low precision, conditional maximization provided a more balanced accuracy. They examined a time frame instead of a single time point because classroom observations of teacher visits include a natural degree of imprecision (Karumbaiah, Borchers, Shou, et al., 2023; Shou et al., 2023).

Reflection Process Supported by Reflecto. During reflection activities, the teacher selects one or more *trends* to explore, analyzes multimodal classroom data to see to what degree the selected trends are present or absent in their own classroom practices, thinks about possible reasons for deviations from desirable trends and/or for adhering to trends that, on reflection, are not desirable, and sets intentions for improved classroom practices. Teachers control their

exploration, with Reflecto offering only minimal suggestions for filtering or sorting the data based on the chosen trends. Reflecto does not itself evaluate which trends are reflected in a teacher's classroom practices. Reflecto was built within CTAT+Tutorshop (Aleven et al., 2025).

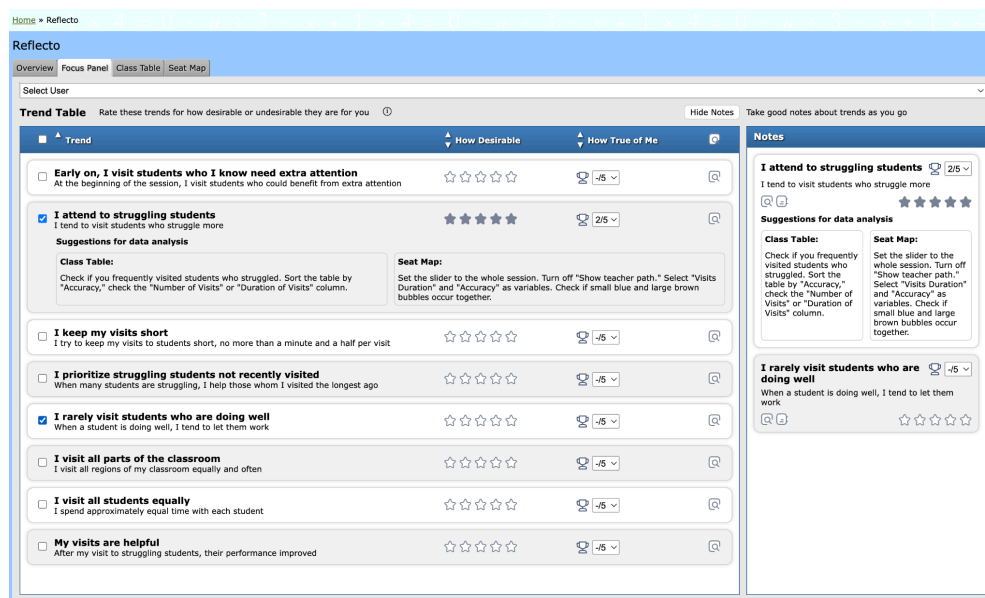


Figure 8.1.1. Focus Panel Showing a List of Trends Derived From Previous User-Centered Research With Teachers

Focus Panel (Figure 8.1.1) Teachers start a session with Reflecto by reviewing a list of eight potential classroom trends (e.g., “I keep my visits short” or “I attend to struggling students”). These trends reflect desirable classroom practices *or* practices to avoid. The trends, informed by user-centered design research with teachers, are meant to capture what teachers might find important as they monitor and help a class of students engaged in work with the tutoring system. Teachers rate the trends in terms of desirability (“How desirable is this?”) and personal alignment (“How true of me is this?”). They select one or two trends to focus on as they explore Reflecto’s data.

One of Reflecto’s two main data displays is a **Class Table** panel (Fig. 8.1.2), which displays analytics about students’ learning, metacognitive, and behavioral states, such as number of problems solved, time worked, accuracy, idle time, misuse time, and consecutive errors. It also displays teacher-student interaction indicators such as visit count and visit duration. Teachers can access detailed visit-level analytics, and can sort the table based on any analytic, for example, to explore correlations between variables (“I attend to struggling students”).

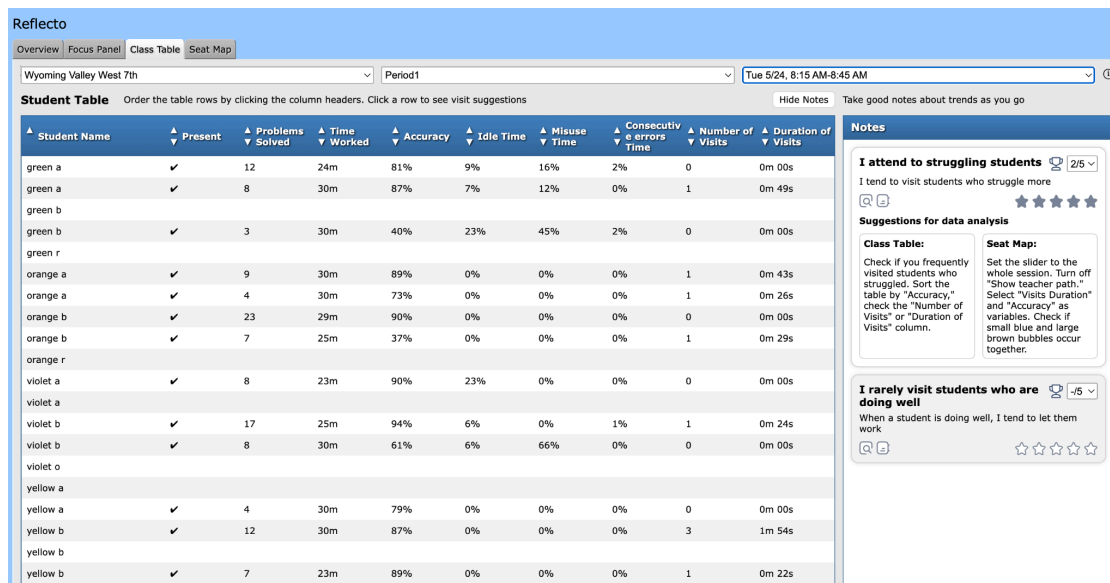


Figure 8.1.2. Reflecto Class Table Showing Students' Learning, Metacognitive, and Behavioral State Analytics, as Well as Teacher Visit Analytics for Each Student

A second data display, the **Seat Map** (Fig. 8.1.3), offers a spatial-temporal visualization of the same analytics, but now superimposed on a seat map. The teacher can select one of the various student analytics and one teacher analytic, toggle visit count and duration on a classroom layout, and track their movement through visit heatmaps or directed path lines, and select specific time intervals of interest. These affordances make the Seat Map a powerful tool for identifying spatial patterns in teaching and learning, such as which areas of the classroom received more teacher attention or which students were most visited.

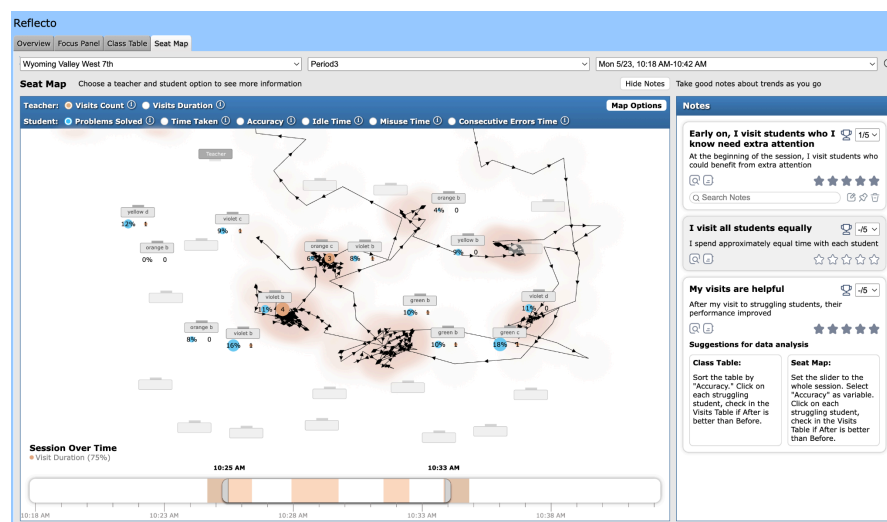


Figure 8.1.3. Reflecto Seat Map Panel Showing Spatial-Temporal Visualizations of Classroom Dynamics, Integrating Student Learning and Teacher Visit Analytics

8.2 User Testing and Prototyping of the Reflection Tool: Reflecto

The design and development of the reflection tool happened mainly during the summer. During the summer when I am away for an internship, research interns (undergrad or graduate students) worked as REU interns under the supervision of the PI (Vincent Aleven) and with mentors of the developers and the research team (including Octav Popescu). Upon completion of my internship during the summer semester in 2024, I returned to the research team and engaged in a series of detailed and focused activities aimed at refining and validating the tool, as well as conducting a classroom feasibility study. These activities included comprehensive functional and stress testing, systematic bug identification and documentation (with severity and impact), and logging bug descriptions and fixing status in a spreadsheet (a method I previously used in testing Pair-Up and related technologies). Findings were regularly communicated to the development team with detailed reports on bugs, fixes, and potential improvements.

8.2.1 Prototyping Activity: Study design, Recruitment, and Data Collection

The prototyping activity's goal is to test with users outside of the research team, how they use the Reflecto tool. The procedure contains 1) prototyping study design, 2) participant recruitment, and 3) data collection. **Prototyping Study Design:** I designed a prototyping study that involved internal team members and external participants, focusing on the usability and functionality of Reflecto.

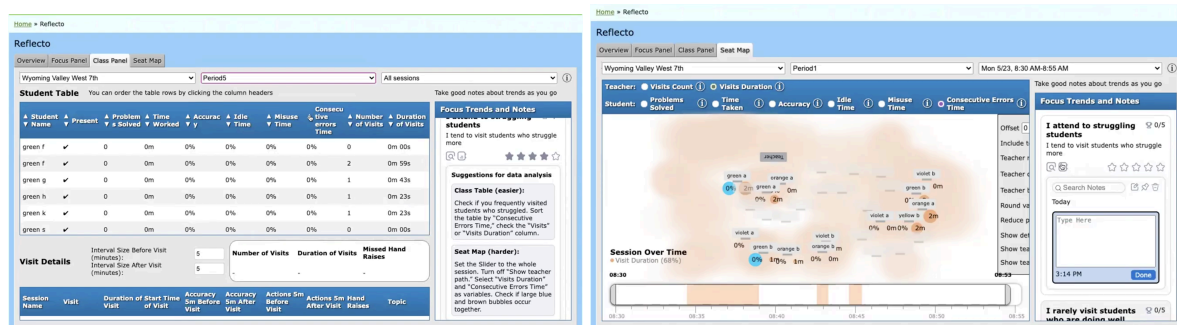


Fig. 8.2.1 Screenshot of Reflecto Class Table and Seat Map Panels during the Prototyping Activities

Tool at the time of study. The Reflecto tool at this stage of the study was close to the final version used in the classroom study in terms of functionality (see Fig. 8.2.1), with the exception of improvements that reduced bugs, enhanced usability, and user experience. These were mostly included: text refinements; changing “Consecutive error” in the data analysis suggestions to “Accuracy”; fixing the data issues existed computation of class table data; improving the time-slider dragging functions; adding an information panel to students’ desks

on the seat map tab; making map options collapsible; and removing certain developer functions (e.g., color, radius, blur).

Specifically, in my prototyping activity protocol, I first asked teachers for permission to record the sessions and collected demographic information about their teaching subjects and experiences. We then asked how they usually structured their classes, whether they had any reflection practices, and how those were carried out, as well as their reflection goals and teaching philosophy.

Next, we introduced three usability practice tasks and four reflection tasks (two individual and two collaborative). Participants began with an individual reflection task, where we asked them to imagine the data shown on the dashboard as analytics from their own classes.

- **Reflection Task 1:** Identify one or two trends you would like to reflect on, and find relevant analytics to determine whether this trend was realized in your classroom teaching.
- **Reflection Task 2:** Formulate intentions on how you would reflect—what teaching behaviors you might want to change—if this were your own data, to improve your classroom practices.

We encouraged participants to take notes during their reflection and to creatively explore the data to gain insights into teaching and learning analytics. After the individual reflection tasks, we invited participants to join a collaborative reflection session. In this session, we played the role of their colleagues from the same subject area. Participants were asked to share their notes from the individual tasks with the co-reflecting teacher. A researcher—either myself or another research intern—played the role of the colleague and asked probing questions, such as why a particular trend was chosen. The co-reflecting teacher then shared personalized feedback, tips, or engaged in further discussion.

We asked participants questions about their regular reflection practices and the tool's usefulness and design, including:

- What were your usual reflection practices?
- Were you able to make sense of the reflection tool and identify its main functions?
- How was your overall experience with the tool? Please rate your experience from 1 to 10 and explain your rating.
- If you were a teacher, what do you think this tool is used for?
- What design feedback or suggestions for improvement do you have?

Participant Recruitment: I recruited a diverse group of teacher participants ($N = 4$), including current and former teachers from three different subject areas, to ensure a broad

range of feedback. I sent out a recruitment message and offered compensation of \$45 for each prototyping session.

Data Collection: I assigned each participant two individual reflection tasks and two collaborative ones. Participants received prompts focusing on their instructional practices, students' learning processes, and the relationship between the two. For example, one prompt asked: *"Please use the reflection tool to reflect on your struggling students' learning process and your interaction with them."* I arranged opportunities for both individual and collaborative reflection to take place.

The student and position data used in the tool during the study were anonymized datasets collected in a previous research study by other team members. Since these data did not represent the actual students of the participants, one limitation was that teachers were not reflecting on their own class data. This might have influenced how they interacted with the tool, as their reflections could differ if real classroom data were used.

I collected feedback from participants to capture their experiences, satisfaction levels, and suggestions for improving the tool. I also used Zoom to record the videos for reflection sessions.

8.2.2 Prototyping Activity: Results and Discussion

Demographics: The four teacher participants, with T1 being the only former teacher, and T2, T3, T4 being the current teachers. T1 taught social studies for 2 years, T2 taught 9 - 12 grade math and chemistry for 10 years, T3 taught math for more than 10 years, and T4 taught high school math for 9 years.

Teaching and reflection practice: For reflection practice, most of them have irregular and informal reflection on how they feel about the teaching process, and whether the students understand the materials. T2 mentioned she received a formal teacher evaluation once a semester, required by the school: *"I think it's just not done any way on paper, but it's more just on myself, could be at the end of the day, how I feel and evaluate myself, or end of the week, or at the end of the chapter. So sometimes I look at tests for assessment students to see if they learn the content. Did I cover the materials well enough that they understood it, but when it comes to anything formal, not necessarily anything formal other than my teacher's evaluation that we do once a semester"*. Similarly, T3 and T4 both mentioned that their way of instruction is a combination of lecturing and having students do math problem-solving. T3 said her usual way of instruction is they lectured the lesson content first, then she'd walk through the problem together with the class, and then students do math problems on their own; while T4 usually did the other way around. She mentioned that their current reflection method mostly **relied on memory**.

Performance in reflection tasks using the tool: After guidance and scaffolding practice tasks as part of the onboarding process, most of the time, the users completed the reflection tasks relatively smoothly without major setbacks.

There were different feedback and actions among the teacher participants:

T1 read through each trend and found the last trend, *“My visits are helpful,”* to be unclear—especially regarding how the system determined whether a visit was helpful or not. She also had questions about the differences between *“how desirable”* and *“how true of me.”* T1 mentioned not having the time to sit down and analyze all of the classroom data by herself, and explained that one scenario where she would have used the tool was when she could select one or two trends that were very important to the classroom and check whether her intuition was correct. She wanted the tool to do some basic analysis for her.

T2 selected Trend 1, *“I visit all students equally,”* rated its desirability to be 5/5 and *how true of me* to be 3/5, as well as Trend 2, *“I attend to struggling students,”* which she rated 4 out of 5 for desirability and 3/5 for *how true of me*. T2 mostly followed the data analysis suggestions, sorted by the *number of visits* in the class table, and in the Seat Map panel she selected a short period of time in the time slider and interpreted the teacher trajectory. She wondered if the time she walked corresponded to circumstances when students raised their hands. From the trajectory, she believed she had walked around the class and mentioned that she liked seeing her path and the accuracy analytics together (e.g., possibly to check for correlations). She used this panel to confirm that she had not favored one side of the classroom over the other, saying, *“I want my students to know that I care for them equally, that their learning is important to me.”* Her feedback was that she believed seeing accuracy as bubbles was useful, and that she preferred the Class Panel to the Seat Map Panel, since she liked the chart with a list of information. She reported that she liked using the teacher path to check whether she had visited all parts of the classroom, even though it could look slightly visually cluttered.

T3 did not notice the data analysis suggestions at first. She noted that *“some students who were doing well still had multiple visits.”*

T4 reflected on all the trends and gave them ratings. She found that challenges arose when data for the column *“consecutive error”* was consistently zero, which limited the usefulness of the suggestion. She described her reflection actions with this panel: *“I moved the sliding bar from the 8:30 to 8:35 mark to show the first five minutes, and then at that time, I saw that the blue (bubble) was only next to an orange in three different places. There were more blue dots, but there were only three orange dots, and they were with the green A student, the yellow B student, and the orange A student. And there was a number in the blue dot that was,*

I guess, the student and how much they had worked. And then there was a different number in the orange dot, which I thought was how many visits I had made or how many minutes there were at that point.”

Intentions formalized. In terms of intentions, T1 proposed it might have been a good idea to pin a list of students’ names who needed extra attention for next time, and mentioned it might have been good if she could *“create categories for my notes”* or have had a place to write *“general, not trend-specific notes.”*

T2’s intention was very concrete, data-driven, and student-centered. She looked at students with zero visits and said that she might have wanted to pay more attention next time. She also realized that visit duration time was usually too short, and that all visits were less than 1 minute. She mentioned she would have wanted to prolong the visits. She realized that she had not visited some students with very low accuracy, which she planned to keep in mind in the next class, for example, *“orange b.”* She, in general, also mentioned wanting to pay more attention to students with zero problems solved, low accuracy, and with no visits last time.

T4 mentioned she wanted to prioritize seating arrangements that supported focus and minimized distractions, avoiding seats near friends or windows. Evaluating whether the current teacher path was efficient or if classroom visits could have been optimized.

Feedback & Places for Tool Improvement. When using seat map panels, T4 thought it was a busy panel with lots of information—mentioning that *“There is so much stuff,”* and *“I might consider several things, like looking at this seat map again without the teacher path, which was a little confusing for me, it was visually cluttered.”*

T2 provided some tool-related feedback, for example, she considered it could have been helpful to show the effect of teacher interactions on student outcome, such as improved accuracy or reduced errors after visits (which actually the tool had a way of enabling teachers, but this feedback showed that it was not obvious/underused to this user, and could have considered a more visually engaging/prominent representation). She also showed interest in knowing real-time alerts on student misuse or idle behavior: *“If there’s like right away, like having or knowing that someone is idle, like I can go to them right away.”*

T3 offered feedback for collaborative reflection: she thought it was very easy to collaborate and talk with one colleague of hers (who was teaching social science, a different subject from hers) since they taught the same students and their offices were located near each other, showing that she valued student knowledge and accessibility when considering collaborative reflection. T4 explicitly mentioned that she did not want to add additional trends and thought the current set of trends was excellent. Both T2 and T4 mentioned that the tool *“Was challenging to use at first, but it could become a valuable tool with proper training.”*

Interestingly, T4 mentioned the benefit of combining the use of both panels, and that they could have been complementary to each other, indicating the value of MMLA to bring more contextual information: *“I think the most useful would be how we looked at the seat map at the end, where we took out the path, and we really looked at the numbers and the seating area. Because if I just look at the chart, if the names are put in alphabetical order on the class panel table, it doesn’t tell me where they sit, so it might look very good numbers and data. But then if I take what I see on the class panel table and the seat map, and I combine that information, I might see that (actions happened) for some reason.”*

Other teacher opinions. T1 thought one of the good uses of the tool was *“if I had a specific goal for my classroom or something I wanted to improve, track my behavior and student’s behaviors to validate. Want to see whether the students have similar behavior in other teachers’ classrooms.”*

Design implications. We improved the guidance materials and practice tasks to: 1) better explain terminology and definitions. 2) identify that teachers sometimes found the zero consecutive error to be confusing, so we later adapted the data analysis suggestions to use *“Accuracy”* as compared to the previous version, which used *“Consecutive Error.”* Participants noted some other design feedback (not yet incorporated). For example, T4 also suggested that the tool could have included reminders for teachers to engage with students they had not interacted with recently. She mentioned, *“maybe if I didn’t talk with a child for 5 days or more there would be a reminder to talk to that child, even if it’s just to say you did a good job this week.”*

CHAPTER 9 REFLECTO CLASSROOM EXPLORATORY STUDY (MOTIVATION AND METHODOLOGY)

Acknowledgement. The exploratory study leveraged the Reflecto tool, primarily designed and developed by Prof. Vincent Aleven, lab designer Jasmin Kim, and developer Octav Popescu, Sahil Gupte, and other interns. It draws the technical foundation of visit detection and pipeline from work by Conrad Borchers, Peggy Shen, Lucia Fang, Prof. Shamyia Karumbaiah, and her collaborators. The user testing, tool setup, and other tasks were helped by research volunteers, including Mandy Pan, Manqing Yu, Ashley Xu, Maria Ma, Yvonne Zhao, Yikuan Fang, Liyi Xu.

9.1 Goal of Evaluation Study

I conducted an exploratory evaluation study to see whether and how the Reflecto tool helps teachers achieve a more reflective practice in their daily teaching. Aligned with what I proposed in the thesis proposal, I conducted a classroom prototyping study involving both teacher and student participants. The students engaged with an Intelligent Tutoring System (ITS) in class, and post-class, teachers utilized a reflection tool designed to leverage multimodal learning analytics (MMLA). This tool is intended for after-class reflection by providing teachers with noteworthy insights from classroom interactions and student performance data.

This exploratory study employed both qualitative and quantitative methodologies to gather data on the effectiveness of the ITS and the reflection tool. Through this approach, I aimed to understand teachers' reflection process, the impact on their understanding and behaviors, and their viewpoints towards tool design.

9.2 Research Questions

RQ1: How do teachers use the MMLA-based teacher reflection tool during individual and collaborative reflection sessions?

- RQ1.1: What trends do teachers choose to reflect on, and how desirable do they consider these trends?
- RQ1.2: What conclusions do teachers draw about the selected trends?
- RQ1.3: What happened in teachers' collaborative reflection activities?

RQ2: How does Reflecto influence teachers' understanding of their teaching practices and future intentions for interacting with students?

- RQ2.1: How does reflection validate or interact with teachers' existing knowledge about their teaching and students?
- RQ2.2: How do reflection activities bring new knowledge to teachers?

RQ3: How does the use of an MMLA-based reflection tool impact teachers' actual teaching practices?

- RQ3.1: What actions or intentions do teachers set following reflection—what do they aim to keep or change?
- RQ3.2: To what extent do teachers carry out the changes they intended—how do their stated intentions compare with their observed classroom behaviors?

RQ4: What are teachers' perceptions of the MMLA-based reflection tool, and what are the design implications derived from the study?

- RQ4.1: What features do teachers find most helpful in supporting their reflection practices?
- RQ4.2: What aspects of the tool did not work well?
- RQ4.3: What are their views on using the tool regularly over an extended period?
- RQ4.4: How can the tool be improved to better support teacher reflection?

9.3 Study Procedure

9.3.1 Study Overview

Recruitment emails were sent to approximately 20–30 teachers prior to the study. We eventually recruited two teachers from Silver Hollow School (alias): Ms. Davis (T1, alias) and Ms. Jones (T2, alias). T1 was a previous participant in one of my previous studies (Pair-Up, Chapter 5), who introduced us to the possibility that T2 may also be interested. Both voluntarily joined the study after reviewing the protocol through email communication. We obtained approval from their school principal. All teachers' and students' names in this study were aliases for anonymity purposes.

9.3.2 Participants and Site Permissions

Recruitment emails were sent to approximately 20–30 teachers prior to the study. We eventually recruited two teachers from Silver Hollow School (alias): Ms. Davis (T1, alias) and Ms. Jones (T2, alias). T1 was a previous participant in one of my previous studies (Pair-Up, Chapter 5), who introduced to us that T2 may also be interested. Both voluntarily joined the study after reviewing the protocol through email communication. We obtained

approval from their school principal. All teachers and students names in this study were alias for anonymity purposes.

The study was approved by our institution's IRB. Under the Exempt Category 1, given that the study posed no more than minimal risk for students, and with a focus on the teacher analytics and data rather than students' learning data, parent consent was waived under the approved protocol; however, we provided teachers with informational materials that could be distributed to families for transparency and opt-out purposes. Teachers signed informed consent forms prior to participating in the reflection activities.

9.3.3 Classroom and Hardware Setup

Location Sensing System. We used Pozyx Inc.'s ultra-wideband location-sensing technology to track teacher movement in the classroom. Four tripod-mounted anchors were positioned in the corners of each classroom to maximize spatial coverage. We measured the room dimensions, removed obstacles at the corners, and uploaded a classroom floor map into the Pozyx system. Each anchor's (x, y, z) coordinates were manually entered and verified against the digital floor plan.

To improve setup stability, we mounted anchors using sturdy metal brackets and marked the tripod positions with tape to ensure consistent placement. Gateway and anchor power adapters were labeled to prevent configuration errors.

Student Seating Configuration. Each classroom was organized so that desks were aligned and ample walking space was available. Each student's seat was assigned a number and its coordinates were measured using a Pozyx tag. For each seat, coordinates were recorded twice and averaged to improve accuracy. T2's classroom study sessions happened one week after T1's, where we additionally labeled desks with period numbers, student names, and seat IDs to encourage consistent seating. Extension cords were provided to reduce the likelihood of students changing seats to charge devices.

9.3.4 Instructional Context

T1's classes: Three 8th-grade classes participated over four class sessions. Students worked on algebraic equation solving using the Lynnette ITS.

T2's classes: Three 6th-grade classes participated over four sessions. Students completed content covering expressions, decimals, proportions, ratios, and one- and two-step equations.

Each class session lasted approximately 40 minutes and took place during the school's regular math periods. Teachers wore the location tracker during these sessions.

9.3.5 Study Procedure

The study was designed to span two weeks. Each teacher experienced four days of classroom sessions, four collaborative reflection sessions (two led by each teacher), and a concluding interview with an exit survey.

Both (T1 & T2) experienced some common structure: 1) **Days 1–2:** Classroom sessions where three classes (for each teacher) used the ITS (Lynnette). 2) **Days 3–4:** Continued classroom sessions with the same classes using the ITS. 3) **Collaborative Reflection Session:** Two sessions for each teacher, one after the first set of classroom days and one after the second. In the session, one teacher was the main reflecting teacher, and the other was the co-reflecting teacher. 4) **Final activities:** Concluding interview and exit survey on the usefulness of Reflecto features. There are some **additional activities unique to T2:** 1) **Individual data exploration and reflection sessions (2 × 30 min):** One before the first collaborative reflection and one before the second, designed to support familiarity with student data and the Reflecto tool. 2) **After Day 4:** Supporting example and counterexample reflection session.

The study protocol was iteratively refined throughout the research process. Adjustments were made based on pilot testing, teacher availability, and emerging needs during implementation, ensuring the procedure remained flexible and responsive to real classroom contexts. The specific iterations & changes were described in Section 9.4.

9.3.6 Metrics and Data Sources

This study utilized a mixed-methods approach, drawing on six primary data sources to examine how teachers engaged with Reflecto, how they interpreted multimodal classroom data, and how their reflection informed instructional practices.

9.3.6.1 Teacher Reflection Observations and Notes During reflection sessions, teachers interacted with Reflecto, a tool that integrates teacher location data and students' ITS log data into visualizations. Each teacher session was observed and recorded (with permission and consent form from teachers). Recordings included either the screen only or both the teacher and their screen (captured using a tripod or handheld camera).

Teachers were encouraged to use Reflecto's built-in note-taking feature to jot down observations, surprises, or reflections during their sessions. These notes provided insight into what elements teachers paid attention to and how their interpretation evolved over time.

9.3.6.2 Teacher Location Data Teachers wore Pozyx ultra-wideband sensors during all class sessions, which recorded their up-to-the-second movement throughout the room. Using

predefined coordinates for each student seat, we were able to calculate which students were visited, how often, and for how long. A stop-detection algorithm was applied to identify moments when the teacher appeared to engage with a student. Limitations in the precision of this algorithm were noted and considered in the analysis.

9.3.6.3 ITS & Reflecto Log Data Students' interactions with the intelligent tutoring systems were automatically logged via TutorShop and LearnSphere platforms. These logs included timestamped data on correctness, hint use, error patterns, and more. In combination with visualizations in Reflecto, these data helped teachers assess student learning behavior during reflection. ITS data was displayed in Reflecto and used by teachers in their reflection. As for data analysis of the ITS log to answer RQs, we analyzed students' accuracy and visits in part to answer RQ3 about whether teachers' intentions were actually carried out.

In Reflecto, it also captures teacher users' transaction and interaction log with the tool, including notes taken, trends selected, ratings given for the trend, student analytic item selected, etc. Many of these were used to answer RQ1.

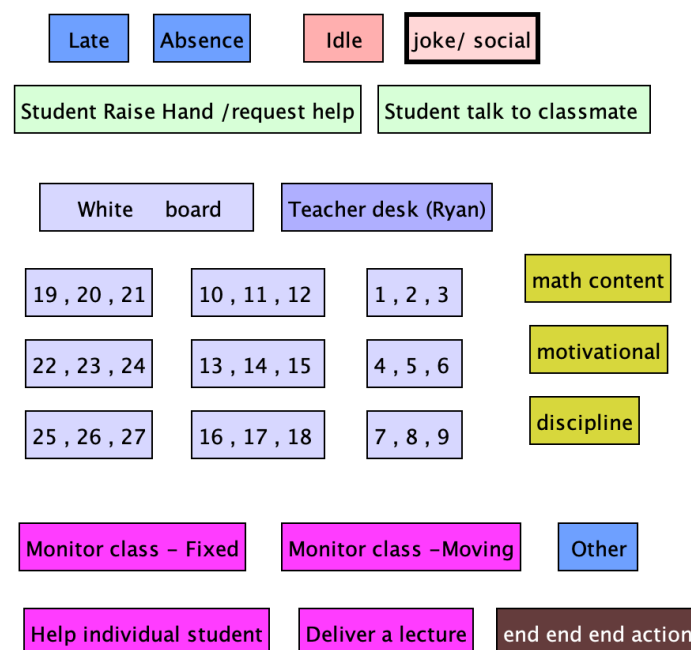


Figure 9.2. Classroom Observation Template in the Look Who's Talking App for T1's Classroom Setting

9.3.6.4 In-Class Observations and Field Notes Research volunteers conducted live observations during all classroom and reflection sessions using a structured framework.

Observation setup. Using the LookWhosTalking app, we created a custom classroom seating chart (e.g., Fig. 7.2) that included numbered student seats (1–27), the teacher’s desk, the whiteboard, and other key furniture. At the start of each class, student attendance was recorded and cross-checked with the teacher. Absent students were marked in the data pipeline to exclude them from Reflecto’s analytics.

Behaviors captured. 1) *Teacher behaviors:* Monitoring (fixed), monitoring (moving), helping individuals (student ID recorded when possible), and delivering whole-class instruction. 2) *Student behaviors:* Late arrival/absence, idle (not on task), joking/socializing, help seeking (hand raise; ID recorded when possible), and peer talk (ID recorded when possible). Because there was typically only one research volunteer present, these behaviors were logged to the extent possible but not exhaustively. Student interactions such as idle, joking, help seeking, or peer talk could occur simultaneously, and we estimate that roughly 60% were successfully captured (see Fig. 7.3). This data source was not systematically analyzed in the present study but was collected as a safety check and supplemental record.

Teacher conversation codes. Interactions were also categorized as math content (subject matter explanation), motivational (encouragement or support), or disciplinary (classroom management/redirection).

	A	B	C	D	E	F	G	H	I	J	K
1	#2025 E/19 10:13:14										
2	time	speaker	speakercomment	x	y	notes	text				
3	02:01:979	Deliver a lecture		190	830	Ryan use the whiteboard and deliver the lecture on the teacher's desk					
4	04:06:386	Deliver a lecture		880	860	Bella went to the 1st and Ryan standing besides the whiteboard					
5	06:02:416	White board		210	350	teacher stand in front of the whiteboard and explain the system					
6	06:09:912	Deliver a lecture		550	950	Bella continued the lecture					
7	08:44:98	Teacher desk (Ryan)		550	350	Teacher back to the desk					
8	10:11:578		3	2	1	590	470		3	teacher helped with the math problem	
9	10:56:505		12	11	10	370	470		11	teacher helped with math	
10	11:29:872	Student Raise Hand Request help			270	210	24 raised the hand				
11	11:43:433		24	23	22	130	570		24	teacher helped the math problem	
12	12:21:112		21	20	19	130	470		21	teacher helped the math problem	
13	13:08:767		24	23	22	130	570		23	teacher helped with the math problem	
14	13:49:655	Student Raise Hand Request help			270	210	17 raised hand				
15	13:58:320		18	17	16	370	670		17	teacher helped with the math problem	
16	14:31:43		9	8	7	590	670		8	teacher helped with the math problem	
17	14:59:250		24	23	22	130			570	23 leaved the seat	
18	15:21:196	Other			850	830	teacher talked with another teacher who came into the class				
19	16:20:368	Idle			490	110	19 looking around				
20	16:58:658	Student talk to classmate			730	210	15 and 14 talked 15 and 10 talked				
21	17:32:465	Student talk to classmate			730	210	15 and 11 talked				
22	17:50:758	end end end action			850	950	teacher stopped talk with another teacher				
23	18:10:259		24	23	22	130	570		24	teacher helped with math problem	
24	18:43:431		24	23	22	130	570		2	3 teacher helped with math problem	
25	19:11:820		24	23	22	130	570		23	leaved the seat	
26	19:23:669		15	14	13	370	570		15	teacher helped with the math	
27	20:01:812	Student talk to classmate			730	210	21 and 24		19 and 23		
28	21:15:710		24	23	22	130	570		23	teacher helped with the math problem	
29	21:53:734		24	23	22	130	570		22	teacher helped with math	
30	22:11:945	Student Raise Hand Request help			270	210	17				
31	22:18:406	Student Raise Hand Request help			270	210	error				
32	22:27:428		15	17	16	370	670		17	teacher helped with the math	
33	22:53:222		9	8	7	590	670		9	teacher asked whether he need help	
34	23:09:537		9	8	7	590	670		8	teacher helped with the math problem	
35	23:40:172		3	2	1	590	470		2	teacher asked are you and then helped with the math	
36	24:08:538	Student talk to classmate			730	210	student started to talk				
37	24:26:980		15	14	13	370	570		15	said stopped talking	
38	24:49:538		12	11	10	370	470		12	asked the teacher and the teacher helped with the math	
39	25:42:257		21	20	19	130	470		20	teacher helped with the math	
40	25:56:832		21	20	19	130	470		19	teacher helped with the math	

Figure 9.3. Sample Data Export Output From the Look Who’s Talking App for T1’s Classroom

In addition to coding behavior, observers took notes of similarities and differences between the two teachers’ practices to support comparative analysis.

9.3.6.5 Teacher Interviews Each teacher participated in a single post-reflection interview following the final reflection session. These semi-structured interviews were designed to gather insights into how Reflecto influenced their thinking, practices, and perceptions.

Interview questions focused on three key areas:

- **Teacher Beliefs and Instructional Practices:**
 - How do they typically support student learning in whole-class and individual interactions?
 - What are their strategies for deciding whom to help, and how do they structure those interactions?
 - Do they usually engage in reflection, and if so, how?
- **Experience with the Reflection Tool (Reflecto):**
 - What aspects of Reflecto were most useful or surprising?
 - Did the analytics surface any new insights or help set future teaching intentions?
 - How did they interpret and use teacher location data and student log data during reflection?
- **Design Feedback and Future Use:**
 - What design suggestions do they have for improving Reflecto?
 - Did they feel Reflecto helped improve their teaching practice?
 - Could they envision using the tool regularly as part of their instructional routine?

The goal of the interview was to understand how Reflecto supported more reflective practice, how it affected teachers' thinking and future intentions, and what feedback they had for improving such a tool. This study did not aim to measure student learning gains as a result of using Reflecto.

9.3.7 Compensation

We documented the exact time that teachers spent outside of class—including individual and collaborative reflection sessions, onboarding, and interviews. Teachers received monetary compensation in the form of an Amazon gift card.

9.4 Iteration to Reflection Activity Protocol and Interview Protocols

Throughout the study, we made iterative refinements to the reflection activity and interview protocols to better support teacher engagement and tool evaluation. The detailed protocols are here:  Updated protocol . The following describes key iterations:

Iteration 1: Adding Accountability in Note-Taking

After the first collaborative reflection, we observed that T1, who does not typically engage in structured reflection, needed additional scaffolding to make their reflection more data-driven. To improve specificity, we revised the note-taking prompt from an open-ended "Please take some notes" to a structured format encouraging three notes per session. We provided both printed and Google Doc versions of this template. To stimulate deeper conversation during collaborative reflection, we added three prompts:

- "What surprised you about this data?"
- "What patterns do you see that I might have missed?"
- "How might our classrooms be different or similar?"

We also updated T1's collaborative reflection protocol to include more discussion prompts and encouraged sharing verbal responses, even if the note template was not completed.

Iteration 2: Adding Individual Data-Exploration Sessions

Despite onboarding, T2 expressed a desire for more time to explore the data during collaborative reflection, but felt constrained by the group setting. To address this, we introduced two 30-minute individual exploration sessions between collaborative reflections, held during lunch. These sessions allowed the teacher to examine student ITS performance, visit durations, and help-seeking patterns.

Iteration 3: Adding Specific Student Instances for Prompting

To foster more detailed recall and discussion, we incorporated student-specific prompts such as:

- "Do you remember why you visited this student?"
- "Do you still agree with this visit?"

Iteration 4: Supporting Analytics Examples and Custom Visit Thresholds

We provided teachers with concrete examples and counterexamples from their classroom analytics to facilitate discussion on each selected trend. Based on observations, we also revised the visit detection algorithm: for T2, who often made brief check-ins, the original 21-second threshold did not adequately capture meaningful interactions. In consultation with prior researchers, we added a configurable threshold to the pipeline and set it to 10 seconds for T2.

Iteration 5: Adding Trend-Based Reflection Tasks

As part of T2's study procedure, we developed and added a reflection activity around trend-based analytics. Teacher 2 was asked to consider: 1) Whether she was satisfied with their behavior in examples supporting the trend, 2) Why counterexamples might have

occurred, and 3) Whether they would act differently upon reflection. This was added mid-study, and was only conducted as an exploration with T2, using data across her 12 class sessions. This activity was added to explore teachers' reactions and reflections when encountering supporting and counter-examples, and to examine whether these experiences foster deeper reflection.

Iteration 6: Enhancing Interview Protocols and Tool Feature Ratings

Following T1's participation, we enhanced the post-reflection interview protocol by:

- Adding questions about collaborative vs. individual reflection preferences
- Introducing a structured tool feature rating: Reflecto's three main panels (Focus Panel, Class Panel, Seat Map) were broken into categorized feature clusters with visual thumbnails. Teachers rated each feature on a 1–5 scale or marked "Have not used".
- Streamlining the interview flow to cover teaching/reflection practices, tool use, feature ratings, and design feedback in a coherent narrative.

9.5 Data Analysis

9.5.1 Data Cleaning & Analysis

All individual and collaborative reflection sessions, as well as the supporting example and counterexample reflection activities, were recorded via Zoom and transcribed using the Otter.ai transcription platform. Transcripts were manually reviewed for accuracy, cleaned for coherence, and segmented at the sentence level to support fine-grained qualitative analysis. Each sentence was treated as a single coding unit and entered as a row in a spreadsheet.

To prepare the data, three researchers collaboratively reviewed the transcripts, correcting errors and refining fragmented speech into coherent sentences. Transcript segmentation followed a sentence-based structure, defined by punctuation (e.g., periods). Each sentence was then independently coded by two researchers using a two-level coding scheme (described below in 9.5.2).

Both coders had qualitative research experience and backgrounds in HCI. Coding was conducted blind to one another's annotations. After initial coding, coders met to review all transcripts line by line. When discrepancies arose, each coder articulated their rationale, and differences were resolved through discussion. To ensure interpretive rigor, any unresolved disagreements were adjudicated by a third researcher. This consensus-based approach helped ensure consistent application of the codes and improved the reliability of the qualitative analysis. If the discussion revealed ambiguities in the coding scheme, these were documented and led to refinements in the scheme. Final consensus codes were recorded in a separate "Resolved Coding" column.

This process was repeated across all seven reflection sessions (four collaborative sessions and three individual sessions, each ranging from 30 to 60 minutes). There was a total of 1931 (T1-1:431, T1-2: 290, T2-1: 596, T2-2: 614) sentences for collaborative sessions, 626 (T2-1: 244+ T2-2: 229+ Supporting/counter examples:153) for individual sessions. We conducted a thematic analysis (Braun & Clarke, 2006) of the coded reflection transcripts to identify patterns in how teachers engaged cognitively with the analytics during reflection. Thematic categories were iteratively refined to capture teachers’ data reasoning strategies, emotional responses (e.g., surprise, uncertainty), and the alignment—or misalignment—between their perceptions and classroom actions.

9.5.2 Two-Level Coding Scheme for Analyzing Teacher Reflections Using Reflecto

We’ve developed a coding scheme with the following content. This coding scheme is designed to capture two complementary dimensions of teacher reflection:

- **Reflection Actions:** What teachers *do* while using the Reflecto tool.
- **Reflection Cognitive Value:** What teachers *get cognitively* from using the tool — how the tool supports reflection, insight, and sense-making.

Each sentence unit was coded with two types of codes: **Reflection Actions** and **Reflection Cognitive Value**. Sentences could receive multiple codes from each level, or none, depending on the presence and relevance of the constructs. All codes were applied using a shared spreadsheet where each sentence was a unit of analysis. Coders were instructed to apply all relevant codes to a sentence and to leave it uncoded if neither level of codes applied.

Reflection Actions. These codes describe the teacher’s interactions, behaviors, or verbalized actions while using Reflecto.

Table 9.1. Code and Definition for Level 1: Observable User Actions With the Tool

Code	Definition
Select Trend	The teacher identifies a trend to reflect on and may explain the rationale behind the selection.
Data Analysis	The teacher discusses how they plan to explore or interpret the data (e.g., filtering, sorting).
Instance Found	The teacher references a specific student, visit, or small cluster, without explicitly linking to a trend.
Supporting Example Found	The teacher cites a case that aligns with the selected trend.

Counterexample Found	The teacher identifies a case that contradicts the selected trend.
Example Explained	The teacher elaborates on the reasoning or context behind a supporting example.
Counterexample Explained	The teacher explains the rationale or conditions behind a counterexample.
Intention Stated	The teacher articulates a future plan or instructional intention based on insights from the tool.
Past Intention	The teacher refers to a prior intention or goal they held before using the tool.

Reflection Cognitive Value. These codes capture how Reflecto supports teacher sensemaking, reflection, and understanding.

Table 9.2. Code and Definition for Level 2: Cognitive and Reflective Gains

Code	Definition
Confirm Knowledge	The teacher affirms that tool-provided data aligns with their prior beliefs or expectations.
Surprise	The teachers express that something in the data was unexpected or contradicted prior beliefs.
Learn Info	The teachers gain new insights about themselves or students, without judging alignment with prior beliefs.
Recall Knowledge	The teachers draw upon memories or prior experiences, without referencing tool data.
Question Hypothesis /	The teachers pose a question or speculates about causes or interpretations of the data.

Iterative Clarifications and Refinements for the Coding Scheme. During the process of refining our coding framework, we found it necessary to iterate on several categories to improve precision and clarity. For example, the original Example Found code was overly broad, so we divided it into three more specific categories. Instance Found was applied when a teacher referred to a particular case, such as a student or group of students, without explicitly linking it to a broader trend. Supporting Example Found was used when the teacher highlighted an instance that aligned with a selected trend, whereas Counterexample Found was applied when the teacher identified an instance that directly contradicted the trend under discussion. Importantly, in individual reflection contexts, these three codes could still be applied even when teachers did not explicitly link their examples to a named trend. In addition, both Supporting Example Found and Counterexample Found could be applied to clusters of students if the group of students was treated as a collective reflective case.

We refined the code definitions: Confirm Knowledge was reserved for moments when teachers explicitly referenced Reflecto's data and affirmed that it supported their prior understanding. In contrast, Recall Knowledge applied when teachers relied solely on memory rather than the tool. Select Trend captured both the act of selecting a trend and any accompanying rationale, but it was coded only once per trend to prevent redundancy. Learn Info included instances where teachers described new insights they gained, and it also encompassed rating behaviors such as evaluating trends on the "How True of Me" scale.

Other code refinements included: Past Intention was used when teachers referred to prior goals or instructional plans, with follow-through evaluations sometimes warranting comparative coding. We expanded the Question code to include teachers' moments of speculation or interpretation, such as generating explanations for observed discrepancies or hypothesizing about causes. Finally, we included a broadly defined Data Analysis code to capture any instance where teachers engaged with Reflecto's analytics or visualizations, even in exploratory ways not necessarily intended to answer a specific question.

9.5.3 Analysis Strategies for research question

This section describes the analysis strategies and data sources for each research question.

RQ1: How do teachers use the MMLA-based teacher reflection tool during individual and collaborative reflection sessions?

For RQ1.1 (What trends do teachers choose to reflect on, and how desirable do they consider these trends?), the data sources were reflection session transcripts and trend selection logs. The analysis strategy involved logging which trend was selected and when (by day/period), recording desirability ratings to capture how aligned the trend was with the teacher's values, analyzing frequencies of selected trends across sessions, and using quotes to understand reasons for trend selection.

For RQ1.2 (What conclusions do teachers draw about the selected trends?), the data sources were reflection transcripts, tool interaction logs, and screen recordings. The analysis strategy included qualitative coding of teachers' analysis steps, extracting "How True of Me" ratings, and analyzing supporting reasoning or evidence cited for each conclusion.

For RQ1.3 (What happened in teachers' collaborative reflection activities?), the data source was reflection transcripts. The transcripts of collaborative reflection sessions were analyzed by extracting the utterances of co-reflecting teachers.

RQ2: How does Reflecto influence teachers' understanding of their teaching practices and future intentions for interacting with students?

For RQ2.1 (How does reflection validate or interact with teachers' existing knowledge about their teaching and students?), the data source was reflection transcripts (Level 2 codes: Confirm Knowledge, Recall Knowledge). The analysis strategy involved counting instances of confirmation (e.g., "I already knew ...") and thematically categorizing the types of knowledge confirmed.

For RQ2.2 (How do reflection activities bring new knowledge to teachers?), the data source was reflection transcripts (Level 2 codes: Learn Info, Surprise, Question). The analysis strategy included counting and coding instances of surprise or discovery, and analyzing themes of new insights (e.g., unnoticed student behavior, interaction patterns).

RQ3: How does the use of an MMLA-based reflection tool impact teachers' actual teaching practices?

For RQ3.1 (What actions or intentions do teachers set following reflection—what do they aim to keep or change?), the data sources were reflection transcripts and post-reflection interviews. The analysis strategy involved coding for stated intentions (e.g., "I will try to ...," "I want to focus more on ...") and analyzing motivations for change using thematic coding.

For RQ3.2 (To what extent do teachers carry out the changes they intended—how do their stated intentions compare with their observed classroom behaviors?), the data sources were teacher visit analytics in Reflecto, classroom observation notes, and video recordings. The analysis strategy included comparing intention statements to actual classroom behavior, creating intention–behavior maps (e.g., intended vs. enacted changes in student visits), and conducting pre–post comparisons of visit behaviors to validate intentions.

RQ4: What are teachers' perceptions of the MMLA-based reflection tool, and what design feedback do they offer?

For RQ4.1 (What features do teachers find most helpful in supporting their reflection practices?), RQ4.2 (What are their views on using the tool regularly over an extended period?), RQ4.3 (What aspects of the tool did not work well?), and RQ4.4 (How can the tool be improved to better support teacher reflection?), the data sources were exit surveys, post-study interviews, and reflection video recordings. The analysis strategy involved identifying the highest- and lowest-rated features, extracting supporting quotes (e.g., "This helped me see ..."), and comparing teachers' expected behavior with their actual behavior in reflection.

CHAPTER 10 REFLECTO CLASSROOM EXPLORATORY STUDY (RESULTS AND DISCUSSION)

Acknowledgement. The data collection and analysis of the classroom study were made possible with the generous help of research volunteers Mandy Pan, Manqing Yu, Ashley Xu, and Maria Ma, under the guidance and supervision of my advisors, Professors Vincent Aleven and Nikol Rummel.

10.1 Results for Research Question 1

The first research question asks: RQ1: How do teachers use the MMLA-based teacher reflection tool during individual and collaborative reflection sessions?

- RQ1.1: What trends do teachers choose to reflect on, and how desirable do they consider these trends?
- RQ1.2: What conclusions do teachers draw about the selected trends?
- RQ1.3: What happened in teachers' collaborative reflection activities?

To answer these questions, we propose an expected model for productive collaborative reflection using Reflecto (10.1.1), and describe what actually happened in the reflection sessions.

10.1.1 Expected Model for Productive Collaborative Reflection Using Reflecto

In this section, I describe the expectations for how teachers can effectively engage in collaborative reflection using the Reflecto tool. This model aligns with Verbert et al.'s model, which consists of four steps that continue iteratively in what the authors call the "learning analytics process model" (Verbert et al., 2013). I now describe Verbert et al.'s model, as well as how it connects with our study and the Reflecto workflow.

The first step in Verbert et al.'s model is i) visualising and presenting data to the user. This can connect with the functions of Reflecto's Class Table and Seat Map Table, which present students' learning analytics and teachers' interaction analytics.

The second step expects that teachers and learners will ii) formulate questions and assess how useful data are for addressing those questions. This connects with how teachers in Reflecto can select pedagogical trends to reflect on, which implicitly formulate the questions "Did I successfully follow this trend (if a desirable trend) or avoid it (if an undesirable trend)?"

The final two steps are the sensemaking process iii) responding to those questions (i.e., generation of new insights), which relate to the data analysis and evaluation of pedagogical trend in Reflecto. The last step iv) perform educationally meaningful actions (e.g., for

learners to change behaviours and for teachers to perform interventions), connects to any behaviors that teachers intend to change in classrooms after reflection.

Specifically, we can decompose the expected model into three components, as follows.

1. Trend Selection

Teachers are expected to choose trends they are interested in examining—especially those they consider important but are unsure they consistently follow (in the case of desirable trends) or successfully avoid (in the case of undesirable ones).

2. Drawing Conclusions from Reflection

After selecting a trend, we expect teachers to engage in the following steps:

Data Analysis: Use the Class Table and Seat Map to explore student-level data relevant to the trends they selected in step 1—either by following built-in analysis suggestions or developing their own strategies. In general, we expect teachers to examine the data displayed in both the Class Table and the Seat Map Table, though specific data analysis methods depends on the trend.

In the Class Table, we expect teachers to explore specific data metrics of interest in detail, and to examine their correlations (e.g., by sorting the metrics in descending or ascending order). In the Seat Map Table, we expect teachers to explore spatial aspects of their classroom practice. Specifically, we expect teachers to interpret the bubble sizes that represent student and teacher analytics, and to consider whether they are correlated in meaningful ways. We also expect teachers to look for broader patterns across the data.

Note-Taking: Record notes in the tool as they engage in reflection, to log down insights and observations from data analytics examined in the tool, to support ongoing or future reflection.

Reasoning Process: When assigning scores in the “How True of Me” column, we expect teachers to base their conclusions on the patterns they observe in the Reflecto tool (following Data Analysis), as well as the relative presence of supporting examples versus counterexamples that align or conflict with the specific trend they selected. For example, if a teacher chooses “*I visit all students equally*” as the trend of interest, a supporting example would be an even distribution of visit time across all students. In contrast, a counterexample would be a case where one student receives no visits or an excessively long visit compared to others. More supporting examples would strengthen confirmation of the trend, while more counterexamples would raise doubts and lead to a lower “How True of Me” score.

Interpreting Counterexamples: Teachers are expected to reflect on any

counterexamples—considering why they occurred and how they might address them—potentially forming new instructional intentions.

3. Collaborative Reflection Behaviors

In our expected model, effective collaborative reflection includes several components: 1) Attending to Peer Data: The co-reflecting teacher helps identify supporting or counterexamples that the main reflector may overlook. 2) Engaging in Dialogue: Partners discuss observations, ask questions, and contribute to the reflective process. 3) Sharing Experiences: Teachers offer personal strategies, insights, and lessons learned from their own classroom contexts. 4) Offering Suggestions: When appropriate, teachers provide constructive suggestions tailored to their colleagues' instructional settings.

10.1.2 Study Results: How teachers engaged in Reflection

I will now review how teachers' actual reflection behavior compares to the expected model (in 10.1.1).

1. Trend Selection (RQ1.1)

Both teachers (T1 and T2) consistently selected trends that are highly desirable. Table 10.1.1 provides details about their selected trends, average desirability, and main reasoning.

Prioritizing Early Support for Students Who Struggle. The most frequently selected trend was **“Early on, I visit students who need extra attention”**, chosen by both T1 and T2 in multiple sessions and rated 5/5 across the board. Both teachers emphasized the importance of getting students on task early—especially those likely to disengage or struggle, e.g., *“Some of them will need that just to get started.”* (T2).

Strategic Use of Short Visits. Both teachers selected **“I keep my visits short”**, rating it 4 or 5 out of 5. They framed short visits as a way to manage limited time and reduce student over-reliance. For example, they said, *“Many of them need my attention”* (T1), and *“If I stay there, they are going to be helpless. Some kids will keep me there the entire class period”* (T2). Short visits were seen as a way to nudge students forward without fostering learned helplessness.

Ensuring Spatial Coverage in the Classroom. **“I visit all parts of the classroom”** was another high-frequency trend, with 5/5 desirability ratings. Both teachers stressed the importance of being physically present throughout the space to keep students engaged. They think it is important to *“Make sure different parts didn't feel neglected”* (T1), citing one possible reason being: *“even just the proximity will keep some kids on task”* (T2). This suggests an awareness that presence—verbal or not—can influence classroom behavior.

Ambivalence Toward Visiting High-Performing Students. In the last session, T2 expressed that they want to explore some new trends that they have not selected before. T2 selected “**I rarely visit students who are doing well**”, but rated it only 3/5, expressing ambivalence. While they preferred to focus on struggling students, there was some hesitation about fully neglecting others. “*If they're working well, I don't want to necessarily hover... that attention can go to the kids who are struggling.*” (T2)

In sum, for trend selection in the collaborative reflection sessions, we found:

Teachers leaned toward high-desirability trends that aligned with their core teaching values. Most selected trends were rated 4 or 5 out of 5, indicating that teachers were motivated to reflect on practices they viewed as important but wanted to verify, refine, or better understand. Teachers repeatedly chose trends focused on **intentional support for struggling students, efficient and equitable classroom movement, and broad spatial coverage.** Trends like “*Early on, I visit students who need extra attention*” and “*I visit all parts of the classroom*” **appeared multiple times** across sessions, indicating that teachers saw these as central, recurring aspects of their practice worth reevaluating, and signal a commitment to equity and classroom presence. Their justifications often acknowledged limited instructional time while emphasizing the need for individualized support. Many framed short visits and strategic movement as ways to manage attention fairly while nudging students toward greater independence.

Trend	Times Selected	Average Desirability	Reasoning
I visit all parts of the classroom	4 times	5/5	Teachers circulated to ensure all students received attention and support, avoiding neglect.
Early on, I visit students who need extra attention	3 times	5/5	Helped students get on task early, especially effective in challenging classes.
I keep my visits short	3 times	4.3/5	Allowed teachers to increase reach and provide quick, targeted support to more students.
I attend to struggling students	1 time	4/5	Essential for keeping struggling students engaged and preventing them from falling behind.
My visits are helpful	1 time	4/5	Visits helped clarify difficult hints.
I rarely visit students who are doing well	1 time	3/5	Prefer to prioritize struggling students over high-performing students.

Table 10.1.1 Pedagogical Trends Teacher Selected in Reflection Sessions & Reasoning

2. How Teachers Came to Conclusions During Reflection (RQ1.2)

Based on the transcript analysis of 4 collaborative reflection session videos, teachers regularly consulted the Class Table and Seat Map when analyzing selected trends. They filtered by metrics like accuracy, visit count, and duration. Their interpretations were often based on visual impressions of the dashboard (e.g., how broadly they moved across the classroom), or memory-triggered cases, as compared to systematic examination of all students. For instance, for “I visit all parts of the classroom,” T1 described the experience as “*like a marathon*” and used that feeling to support a high rating for “how true of me”, sometimes without supporting or counterexamples cited. Teachers often centered their reasoning around a few **salient student examples**, especially those with known academic or behavioral challenges. These cases helped them assess whether a trend felt true. For example, T1 cited Quincy as a counterexample when considering whether she consistently helped struggling students. T2 referenced Janelle, Dominic, and Gabriel to reason about how early visits helped get students started. These examples served as mental anchors for drawing conclusions, though not always balanced with a broader scan of class-wide evidence.

For note-taking, teachers took brief but thoughtful notes in the tool. These often emphasized **pedagogical principles and values**. Examples include: “*Start with these students to ensure they get started.*” (T2, Early Visits); “*Nudge the students in the right direction to push them to do it on their own.*” (T2, Short Visits); “*They want instant gratification and it helps some students and not others.*” (T1, Visit Helpfulness). These notes logged **instructional priorities**, rather than student-specific data-driven observations.

Teachers occasionally noticed students who didn’t align with the trend. However, **reflection on these mismatches was brief** and often not pursued unless prompted. Even when counterexamples were found (e.g., Tristan’s low accuracy despite a visit), they were more likely to be explained away than to trigger revisions of their conclusions or actions. Despite mixed or limited data, teachers often rated trends as highly true (4–5 out of 5). This suggests that self-ratings captured **aspirations** or **alignment with teaching values** (e.g., supporting struggling students, equitable coverage), rather than strict data confirmation.

3. Collaborative Reflection Behaviors (RQ 1.3)

This analysis examines the discourse of two co-reflecting teachers across four reflection sessions, in relation to the following expected collaborative reflection model:

Attending to Peer Data. T1 occasionally viewed, referenced, and questioned T2’s data. For example, “*Ms. Jones, what content did they go to second?*”, “*She has 76% accuracy and 15 questions done... that doesn’t add up.*” T2 lightly referenced T1’s reflection: “*Dave (alias) had really high accuracy, right, earlier?*”

Engaging in Dialogue & Sharing Experiences. Both teachers contribute to light discussion, often limited to affirmations or brief probes. For example, T2 suggested data analysis strategies: *“Isn’t that what you want to be analyzing?”* and encouraged her colleague: *“I felt like on the seat map, Ms. Davis, you need to give yourself enough credit”* (T2). T1 mentioned deep knowledge of the kids: *“We’ve known these kids... you know how they’re going to act.”*, and asked questions about software: *“I think she’s asking, is it how long they were logged in or actually working?”*. T2 shares personal stories frequently and vividly: *“My kids are accustomed to using the hints. So they might use them quicker.”* *“Some kids, if I stand by them, they’re like, ‘Why are you here?’”* T1 also shares more modestly: *“I used to let them listen to music during work, but now they complain.”*

One notable thread of sharing involved comparing Grade 6 and Grade 8 students, with teachers highlighting differences in autonomy, behavioral maturity, instructional needs, and interactions with technology or support tools. T2 emphasized that while eighth graders show greater autonomy, they often lack discipline, exhibit stronger rebellion, and are less compliant with classroom rules, particularly around technology use. *“The autonomy in eighth grade is great, but they have less discipline because they think they know it all... The younger grades are better with following those rules than the older kids are, especially with headphones and staying in the right window”* (T2). T1 contrasted her 8th-grade class with T2’s 6th-grade group in terms of academic composition and peer modeling *“I have all the basic or below basic kids, so Ms. Jones right now has some advanced kids and some general kids”* (T1). T2 further noted that differing instructional histories shaped how students engaged with digital supports such as hints: *“My kids might just be more willing to use [the hints] quicker because they rely on them so heavily... you might see a difference with how my kids use the hints as opposed to hers”* (T2).

Offering Suggestions. This was the least common category for both teachers. Most suggestions were indirect or concerning how to do data analysis rather than what to do in classroom. For example T2 said: *“You can change your ‘Accuracy’ instead of ‘Consecutive Error’ and check it... compare to your visits.”*

In conclusion, for collaborative reflection behaviors, teachers mostly engaged in effective collaboration, generally aligned with our expected models. Teachers shared extensively about their own classroom teaching experiences and engaged in discussion. Teachers seldom offer teaching suggestion to the other teacher, and occasionally reference others’ data.

10.2 Results for Research Question 2

RQ2: How does Reflecto influence teachers’ understanding of their teaching practices and future intentions for interacting with students?

- RQ2.1: How does reflection validate or interact with teachers' existing knowledge about their teaching and students?
- RQ2.2: How do reflection activities bring new knowledge to teachers?

10.2.1 RQ2.1 How does reflection validate teachers' existing knowledge (Confirm Knowledge)

This section describes the affinity diagram results for quotes coded under *Confirm Knowledge*, defined as: *"Reflecto's information or data confirms what teachers already knew about their students or their visits."* These quotes were categorized into three overarching themes: 1) Confirmation of students' performance and learning ability (N = 19), 2) Confirmation of student traits, behaviors, and teacher-student interaction (N = 22), and 3) Confirmation of teacher visits and instructional strategies (N = 26).

1. Confirmation of Student Performance and Learning Ability

In this category, teachers concluded that the student data—such as accuracy, misuse time, and problems solved—matched their prior impressions of student performance and work style. For example, one teacher stated, *"My accuracy probably matches up well with where my kids perform in class for the most part"* (T2). Teachers also made general comments about student work styles and behaviors, such as, *"That's how Charles is for sure. Mikhail is just in his own world normally"* (T2), and *"It doesn't surprise me with Ian"* (T2). Emotional states were also referenced: *"He gets very frustrated"* (T1).

Teacher 1 emphasized that much of what Reflecto showed was unsurprising: *"Honestly, nothing surprises me, just because I'm literally running around this classroom all the time, or at least directing them to be on task"* (T1). They explained this by saying, *"We've known these kids—we have them 84 minutes a day all year. So by this point in time, you know how they're going to act"* (T1).

2. Confirmation of Student Traits, Behaviors, and Interaction with Teachers

This theme captures how teachers confirmed their understanding of students' traits and how they interact with teacher support. For instance, one teacher confirmed certain **students who do need extra help**: *"He's someone that might need a little help to get started"* (T2), and *"So he's someone I try to pay attention to, because he needs the extra support. Now someone like Nick will sometimes help keep him on task—like give him a nudge—rather than from me"* (T2). In contrast, some students were recognized as **resistant to help**: *"Asa does not surprise me. Asa is the one I had to go back and forth. I would stand here, he'd turn this way. I would stand here, he'd turn the other way"* (T2). As another example: *"Monica's always quiet. I wouldn't visit her because she's a behavior issue"* (T2).

Teachers also reflected on how well students comprehend teachers' instructions and their question-asking strategies. One noted, *"His questions are always valid questions, good questions. But he definitely will ask five questions before he gets started, just to make sure he's in the right direction. As opposed to someone like Zion, who probably needs to ask questions and never does"* (T2). T2 added, *"My kids are pretty good with raising their hand if they need help"* (T2). Teachers also confirmed that sometimes students were distracted or engaged in off-task behavior: *"He'll just choose and not follow them so he could finish a game or do something else"* (T2).

3. Confirmation of Teacher Visits and Instructional Strategies

The third theme captures how teachers—primarily Teacher 2—confirmed their knowledge of their own classroom movements and instructional strategies, especially in terms of student visits. Teacher 2 confirmed that she moved around the classroom and covered all areas: *"But it looks like I was, you know, I did a fairly decent job of being everywhere in the room,"* and *"I feel like the browns (circles) are everywhere. I mean, the heat map. So I feel like it looks like, yeah, I'm visiting everywhere."*

She also confirmed providing support to specific students: *"You can see, I feel like my heavy foot pattern here—that's someone who would need help getting started for sure."* T2 commented that the number of visits didn't vary too much: *"Other than him, there's not a huge range in visits. For the most part... 14 is my highest, two being my lowest... I guess it's not even an outlier because I have some fours, fives... oh, there's a three in there. So I don't know. I try to visit them all."* She further reflected on whether there are outliers in visits: *"I'd say... I don't know that statement is super accurate in this class. It seems like I visit them a fairly decent amount—everybody. There's a few well under, but there's no one crazy above, I feel, as an outlier data."*

T2 believed that her visits were helpful, saying, *"And my visits are helpful. I hope they are. Let's say it's a four (how desirable). A lot of the times we could see the accuracy improve after the visits. So I think that supports that."* She also confirmed her strategy of keeping visits short and giving students quick nudges: *"I'm going to look at my period 5. Keep my visits short. Nobody really had overly long visits. I mean, this one had seven—Jay had seven—but that was two minutes total. So it's kind of touch base with her, nudge her in the right direction, then move away."* Finally, she mentioned that simply being physically close to students could be supportive: *"Just being close will help."*

In sum, teachers sometimes used Reflecto to confirm some of their existing knowledge about students' academic performance, traits, and classroom behaviors, as well as to validate their own instructional strategies and movement patterns. These confirmations ranged from

matching data with known student traits, to affirming visit frequency, impact, and the effectiveness of short, targeted support.

10.2.2 (RQ 2.1) How does reflection validate teachers' existing knowledge (Recall Knowledge)

This section describes the quotes coded under *Recall Knowledge* for Level 2, defined as: “teachers mentioned and recalled memory about what they remember happening in class, or their knowledge of the students.” There is a total of 306 utterances under this code, which primarily fell into four categories: 1) Teacher recall of student performance and classroom presence (N = 46); 2) Teacher recall of students’ help-seeking behaviors, learning habits and traits (N = 87); 3) Teacher recall of classroom events, discipline, and atmosphere (N = 77); 4) Teacher recall of instructional strategy, teaching behaviors, and philosophy (N = 99)

This section is longer than other subsections of 10.2 because *Recall Knowledge* contains the most quotes. It’s worth noting that in themes 1 and 2, there isn’t always a clear-cut boundary between teachers’ reflections on student performance, classroom behavior, and personal traits—teachers often connect and comment on these aspects together.

Theme 1: Teacher Recall of Student Performance and Classroom Presence (N = 46)

This category primarily includes instances when teachers, based on Reflecto data, recalled whether certain students were present on a given day or commented on students’ classroom behaviors and performance based on prior knowledge.

For example, Teacher 1 **questioned the accuracy of data when she believed a student was absent**: “*This is interesting because I’m pretty sure that Zack was not in class. Yeah, I believe that’s the day I kicked him out. I don’t even think he had a chance to even sign on. Unless he just signed on and then never got himself out*” (T1). Teacher 1’s reflections on **student performance** were generally brief or broad, such as: “*He actually worked so hard*” (T1). By contrast, Teacher 2 offered more detailed and specific reflections tied to individual students’ performance. She combined Reflecto data with a recollection of classroom activity. For example, she discussed her surprise at a student’s high misuse of time despite recalling that he had completed a significant amount of work: “*Like John—he has 26% of misused time overall... That surprises me for him. He’s one that got through. I think maybe all of them—all the question sets. He got through to fractions, decimals, and percents. He might have been on the expressions. So like he got through a decent amount of questions. So to have a 26% misuse time surprises me*” (T2). This quote illustrates how Teacher 2 drew on detailed classroom memories—including content coverage and pacing—when interpreting analytics data, in contrast to the more general commentary from Teacher 1.

Theme 2: Teacher Recall of students' help-seeking behaviors, learning habits, and traits (N = 87)

Teachers often recalled student traits, personalities, and learning styles during their reflections. These utterances can be grouped into three subcategories: **(1) help-seeking behaviors, (2) learning habits and behaviors, and (3) personality traits.**

Regarding **help-seeking**, teachers described students who avoid or reject assistance. For example, T2 said, *"He definitely needs less help and he's also not willing to ask for help. Like he's kind of quiet. Yeah, there are definitely students in class who are just on their own."* (T2) She noted some students may be perplexed or not at ease with teachers being around: *"There are some kids that when they're working, when I walk around, if I stand by them, they're like, 'Why are you here? What are you doing?' So the visits don't benefit them if they're doing okay."* (T2). T1 echoed that students differ in their independence, and differentiated different types of help: *"Some are warranted, some are not warranted. Some kids will ask for help no matter what"* (T1). Both teachers mentioned it's important to keep students accountable: *"Because there were some students that just want you to sit next to them and tell you exactly what to do"* (T1), and students differ in how thoughtful their questions are: *"They're better at asking, [...], actually have intelligent questions instead of 'Do this problem for me'"* (T2). T2 reflected in details how a student can be over-relying on teacher's help: *"So Tyler needed more support, but if I spend time with him, he's going to want me there the entire period to do his work for him. He doesn't push himself ever to work on his own. He just sort of gives up, calls it a day, or I have to hold his hand through it"* (T2). Teachers thought this help-seeking behaviors of students may change overtime: *"He's getting better at being like, 'Oh, help,' or 'I need this'"* (T2).

Teachers also commented on students' **learning habits**. T1 noted that many avoided or misunderstood hints: *"A lot of them didn't want to read the hints, and if they read them, they weren't understanding them"* (T1). T2 discussed whether students show their work: *"I have kids that'll turn in a test or a quiz and they have everything correct, but there's no work, for questions they wouldn't solve mentally typically - so they're just too lazy to write the work down. There's kids who fill the page with work and none of it makes sense. I think we just learn our students: whether they're being lazy or whether they genuinely know something"* (T2). She also reflected on students' differentiated capacity to follow instructions: *"Typically, there are students - even if I gave them five-step directions [verbally]—they'd be able to follow them. There are students that would have to be one to two steps at most"* (T2). T1 briefly commented on student **efficiency**: *"And Jorden's really smart if he actually works... yeah... um... if he's not sleeping"* (T1). T2 offered a detailed observation: *"So John [...] he'll be three-fourths of the way done with the assignment before the other kids are like done with the first quarter. He's really smart. He works hard. He's efficient. He's efficient because he*

wants to be done. Although he gets frustrated when he's not accurate as well. So he's got a good combination of competitive to be done, efficiency to have his own time, but he does care if he doesn't do well" (T2).

Finally, teachers recalled **general personality and behavioral traits**. T1 mentioned students being in **low energy**: *"Mike is always—oh my gosh—yeah, he's always sleeping. I am always, 'Do we sleep like it's a late night here, bud?'"* (T1) T2 added: *"Mitchell is just in his own world normally. But he's the one that once he starts working, when he's working, it's kind of like, let him go, don't bother him"* (T2). Teachers recalled students' **motivation**: *"Most of them are not going out to seek attention to get started. They'd rather sit there with a blank stare on their face. Not very self-motivated individuals"* (T2). Teachers also recalled cases for students with **special education needs**: *"A lot of them are special ed students, and so they struggle with reading. So if there's a lot of directions, a lot of language in there, it'll help to stop in with them just to make sure that they understand what's actually being asked of them. Carly's one—she's not special ed, but she's just all over the place all the time. She comes to school to be social, and so she needs a little bit more redirection just to stay focused"* (T2).

Theme 3: Teacher Recall of Instructional Strategy, Teaching Behaviors, and Philosophy (N = 99)

The most common sub-theme in this category relates to how teachers recalled and reflected on their interactions with students—especially their visiting practices—either in general or in reference to specific visits shown in Reflecto data.

Teachers mentioned they employed a **differentiated visit strategy for different times in class and for different students**. Both teachers mentioned visiting students early if they needed support. T2 said, *"So I would say early on, I visit students I know need extra attention. Some of them will need that just to get started"* (T2). T1 shared a similar principle: *"With my kids—probably T2 has the same issue—is that if you don't get them started, they'll sit there and wait"* (T1). T1 recalled helping struggling students: *"So I definitely did attend to the struggling students, which—we were on the struggle bus—because they just do not listen to directions or watch demonstrations"* (T1). T2 added emphasis for the importance of understanding why students struggle, not just attending to them: *"So not that attending to them is not important—like, it's important—but why they're struggling, sometimes determining whether they're struggling versus being lazy"* (T2). She also noted a relevant challenge is that struggling students may not reach out: *"Sometimes prioritizing the struggling students isn't quite as easy because there're kids that maybe are raising their hands that want help that can probably do it, versus the ones that are struggling and just sitting in silence. So it's not that you don't prioritize them. It's just getting to them in that*

priority order sometimes is difficult” (T2). For students doing well, T2 preferred to give them space: “If they’re working well, I don’t want to necessarily hover over them and let them go. And then that attention can go to the kids who are struggling. But I don’t want to ignore them either” (T2).

In teacher visits, teachers strongly emphasized proximity, keeping visits short, and not visiting students equally. Both emphasized **proximity as a strategy to increase student focus**: *“I feel like proximity with kids just helps in general. Like when you’re close, they’re going to be more on task” (T1). Based on the reflection, T2 recalled the specific, differentiated visiting strategy over time in class: “I feel like when it first starts and they’re getting started, I loop around specifically to make sure that I’m close to everybody at some point. That way, if they have a question, they’ll ask me when I’m there. And then I sort of navigate around by where the hands go. And then once the hands are not going up or down, you can see who looks to maybe be off task or doesn’t look as engaged as they should be” (T2). This represents an interesting reflection that may not be made possible without analytics shown on Reflecto’s data. T2 strongly emphasized keeping visits short, “So early on, I think that’s definitely—so I try to keep my visits short [...] So my visits are shorter just to try to nudge them in the right direction and then see what they can do on their own” (T2). Her reasoning ties to educational psychology: “Because it’s sort of like a learned helplessness—if I stay there, they’re going to be helpless and not be able to complete what they need” (T2). She valued this strategy so much that she increased the desirability rating: “I actually went and changed my ‘keep my visit short’ to a five from a four” (T2). She noted that one visit could benefit nearby students too: “So there were times that even if they weren’t on the same question, what I was helping would help everybody that was sort of in that area” (T2). T2 firmly disagreed with visiting all students equally: “I don’t think that needs to happen by any means. There’s going to always be kids that need more / less attention than others” (T2).*

Reflection based on Reflecto visit data can help teachers recall the specific reasons why certain visits happened. Such scenarios can give teachers new knowledge about students, based on MMLA data support. T2 recalled helping a student log in: *“In the beginning, it looks like when class first started, I spent almost two and a half minutes with her. She’s someone that might have tried getting logged in on the wrong site. We got her in, and then I kind of hung around with her to make sure she was understanding what we were doing. She does struggle in class, although her accuracy after the visit did go down” (T2).*

For pedagogical strategies, T1 valued students’ accountability through peer learning and having them show their work; and T2 drew a visual demo on students’ table as well as reinforcing students to explain. : T1 recalled one teaching strategy that she encouraged in class - **peer learning**: *“Some of them will get certain skills faster than others. And then*

they're excited that they got it, so then they'll try to help their peer. So that's good" (T1). She also highly valued students' accountability: *"Keep the kids accountable. I think I told you this—I didn't really have time during the class to even look at this program, so there's going to be some accountability or an expectation piece with it. Usually I'll do one and then hope that they can get it"* (T1). She described a strategy she created: *"So I made a rule—I'm like, 'If it's not written on paper and I don't see any hints, I'm not coming to you. Yeah, I'll bypass you'"* (T1). T2 recalled reinforcing students' learning through explaining how the answers are reached: *"[...]We reinforce them explaining the how and the why. So those expectations are asked of them often. Like, 'Okay, you got 42, but how did you get 42?' And we don't typically say, 'Okay, well, we added 20 plus 22.' We say, 'We added the total number of boys at the party to the total number of girls at the party to get the total number of people at the party.'"* (T2). She sometimes uses **visual demonstrations** by drawing on the table: *"I would talk with one kid, we'd start to draw on the table, and I would go through setting up a proportion. And then the other kids would stop and look—the same concept of two boxes, two boxes, one box, two boxes—and how that transcribed into what we would look at as a proportion"* (T2).

Theme 4: Teacher Recall of Classroom Events, Discipline, and Atmosphere (N = 77)

This theme includes teachers' reflections on classroom management, the atmosphere during instruction, variability in student productivity across different content, and logistical or instructional challenges.

Some reflections recalled classroom events based on the Reflecto data, especially T2's comments on how students' productivity varied with content. For example: *"Again, the decline could be that she had cruised through the portions—the problem sets that they were most familiar with—and now is on sort of newer material"* (T2). She reflected on how content familiarity can make a difference on students' productivity: *"It seems like they get more productive and then a little less productive when they get to the different content at the end. So the newer content would have been the expressions and the equations, which I thought we would have had like a weekend on. They were really just briefly exposed to it, not really taught the content"* (T2).

More in this category are not necessarily based on the Reflecto data but are more on logistical issues. For example, T1 described preparation steps to get students logged in: *"I did make QR codes for the software and then just put it around the room so they could scan it, and I put all their usernames and passwords on a note card"* (T1). T1 reflected the effectiveness of doing a demo: *"We just gave them an example and showed them. It seemed to help. Some of them still weren't listening, but it definitely lessened the confusion"* (T1). T1 also recalled students reacting to the presence of researchers during the classroom study: *"Why are they*

here?’ ‘Who are these people?’ ‘Why do we have to do this?’” (T1). T1 recalled logistical and technical barriers like charging iPads: “Another issue was like, their iPads—they don’t have them, they’re not charged” (T1). She added that students needed extension cords to use their devices while seated: “So like, one day I borrowed an extension cord because they measure the seats, so like, computers could be at their seats” (T1).

Teachers recalled some additional challenges in software misalignment with the curriculum and translation issues. Both teachers mentioned challenges that came from misalignment between their usual instructional notations and the software interface. T1 said: “Because they’re not used to learning that way. And they’re also very needy. So they need you to hold their hand and walk them through it. And it’s very difficult. I said at the end of the day, you can do it any way you want—from that, there’s only one answer” (T1). She also expressed concerns that middle school students struggle with recognizing different forms of the same problem: “Middle school kids especially are creatures of habit” (T1). T1 recalled challenges with classroom discipline, particularly in one difficult class: “Six—I was like, I’m surprised my hair is not all gray right now because it was just a lot to get them to like...” (T1). T2 added that technical challenges, such as translation issues for ESL students, posed barriers: “He’s my ESL student, [...] I think maybe because the translator kept going on and off for him. Because normally, if I give them things in Spanish, he does a better job focusing.” (T2)

In this section 8.2.2, teachers recalled whether student performance and presence aligned with their memory, made frequent comments on students’ traits, behaviors, and learning style. Teachers also recalled the general classroom atmosphere, discipline and in-class events. As for their own behaviors, the teacher recalled a lot about their visits, including principles and rationale behind certain visits, and memories of why certain visits happened. Teachers also recalled in general about classroom setup, preparation, and events, technical issues that may not be necessarily triggered by viewing Reflecto data.

10.2.3 (RQ 2.2) How does reflection infer new knowledge to teachers (Learn Information)

In this section, we (3 researchers) analyzed the quotes coded under *Learn Info* using affinity diagramming. *Learn Info* is defined as: “teachers learned some information from Reflecto about themselves or their students, but did not mention whether it confirmed or rejected what they already knew.

We roughly identified three types of teacher behaviors under *Learn Info*: 1) Teachers examined student performance and simply noted metrics without interpretation (N = 22); 2) Teachers used student data to make interpretations or identify patterns (N = 21); 3) Teachers learned about their interactions with students, such as visits (N = 16).

The first type of *Learn Info* quotes was brief comments where teachers examined student performance and simply stated the numbers of **certain metrics without further reflection**. For example, one teacher said, “*Jack got a 90% accuracy*” (T1).

The second category included cases where teachers analyzed student data to identify trends or draw conclusions. For example, one teacher noted, “*Her accuracy and her misuse of time are both high on that day*” (T2), and “*Carly has 51 questions, a decent accuracy. She only has five visits*” (T2). Instead of focusing on a single session, teachers sometimes used data **across different time periods** to assess progress over time: “*But she’s at 60 on Monday. On Tuesday, she’s at 49. So that declined*” (T2). In other cases, teachers made observations about the **whole class** rather than individual students: “*Yeah, this class has a lot of high misuse of time*” (T2); “*So yeah, I mean, from this class, it looks like I get everywhere a decent amount*” (T2).

In the third category, teachers learned about **their own interactions with students**. For instance, they noted how often they visited certain students: “*London had four visits*” (T2); “*So Eric did have... three minutes in visits. He was one that was getting to the expressions work*” (T2). They also reflected on the nature or reasoning behind those visits: “*So those were all short visits. Same with Amy. Now they sit really close to each other, so I might have been standing in between them—sort of back-to-back*” (T2). In some cases, teachers found that the data validated their teaching strategies. For instance, where T2 prefers short visits: “*Here you can see that all the visits were pretty short. Emea had eight visits, but there were only three minutes total in length*” (T2). In one interesting case, a teacher realized her actions **contradicted** her intended strategy of keeping visits short: “*It’s interesting to see my pattern in my classroom. Avery’s a question-asker. She’s someone that—if I spent a lot of time—she would want her hand held the entire process. So it’s got to be quick. But it looks like I did spend a decent amount of time. She has five visits*” (T2). This instance illustrates how Reflecto helped teachers recognize when their enacted practices diverged from their intended strategies.

In sum, teachers in general used Reflecto to learn new information about student performance, identify patterns across data, and reflect on their own interactions with students.

10.2.4 (RQ 2.2) How does reflection bring new knowledge to teachers (Question & Hypothesis)

This section analyzes teacher reflections categorized under the code *Question*, defined as instances in which *teachers expressed uncertainty about the reasons behind teacher or student actions, questioned how or why analytics were displayed in certain ways, or offered hypotheses regarding what the data might represent*. A total of 48 quotes were coded in this category, falling into two major themes: (1) questions in the sense-making process and

hypothesizing reasons behind analytics (20 quotes) and (2) questions and hypotheses related to the Reflecto tool (28 quotes).

Questions in the sense-making process and hypothesizing reasons behind analytics. One of the most interesting categories of questions is when teachers form **hypotheses about potential rooted reasons behind analytics** displayed in their reflection. During the reflection sessions, teachers form **hypotheses about student behavior**, including possible cheating, misuse of time, or off-task activity. One teacher reflected, *“This class has a lot of high misuse of time... it wouldn't surprise me if they just took the easy way and clicked their way through”* (T2); *“He probably cheated; he probably had his phone somewhere”* (T1). Teachers frequently reflected more deeply by forming **hypotheses about the causes of student struggles**. A key theme was the **challenge posed by language complexity**, rather than math content itself. For instance, one teacher noted, *“My guess would be the kids that are struggling might struggle more with the words than the actual math context”* (T1). Similarly, another observed, *“She's someone that would struggle with the language part of it—probably not the math content”* (T2). A teacher hypothesized that **stress and physical activity** may have played a role in student performance. One teacher said of a student, *“Mary, she doesn't handle stress or anxiety well... Her misuse of time is high and her accuracy is low... But depending on the day, what's going on outside of math class could really throw her off. And this was also the class we had after dodgeball...”* (T2). Other potential reasons the teacher hypothesized include language barriers for ESL students and students failing to connect classroom instruction with the ITS environment. These hypotheses represent a valuable source of reflection that could not be achieved without these contextual MMLA data, which support teachers to probe deeper about possible reasons behind students' performances and teachers' visit analytics.

Interestingly, teachers used Reflecto analytics as a prompt to question and reflect on their **own instructional behaviors**, especially in relation to patterns of movement and student visits. In several cases, teachers questioned why they appeared to spend a higher-than-expected amount of time in certain areas of the classroom. One teacher asked, *“Who is here [top right]? Why did I stop here all the time? I know it looks empty, but I don't know”* (T2). The Seat Map panel captured and offered contextual seating information that allow teachers to reflect on how **physical constraints of the classroom** affected their movement: *“She would ask questions when I'd be over here, like pinned against the backside table. So going behind her in that area is not quite as convenient”* (T2). Teachers also **recalled the likely content during their interactions** with students based on the Reflecto display of lengthy or frequent visits. One teacher explained, *“There are big [visit] circles around this table[...]. I probably gathered their attention and said, ‘This is what it looks like... These are the proportions... This is how we make a proportion and solve it.’ That's probably what those circles are about—explaining”* (T2). Teachers posed **clarifying questions to their**

co-reflecting partner during the collaborative sessions, e.g., to verify class composition or interpret data. For example, one teacher asked, *“Ms. Jones, in sixth grade, they’re still mixed, right?... I always have all the basic or below basic kids”* (T1). Another asked their partner to confirm their sense of classroom coverage: *“I don’t know... Ms. Davis, what do you think? I hit most of the classroom?”* (T2).

Questions Related to the Reflecto Tool. Many teachers raised questions about the **accuracy and interpretation of the data** presented by Reflecto. For instance, one teacher noticed that the system displayed identical working times for all students in a class period and wondered whether that was accurate, saying *“I did notice they were all 55 minutes. I don’t think that they all probably have exactly the same—there should be a little bit of variation”* (T2). Another teacher questioned whether a displayed misuse time exceeding 100% reflected a flaw in the software, stating *“Is that a program flaw... like with the 105%? Like there has to be something wrong with that”* (T1). Teachers expressed **confusion about student identities** when encountering duplicate names. One teacher noted, *“Well, there are two Jackies (alias). I am confused,”* and commented, *“It’s hard to tell which students are which”* (both T1). In another case, a teacher was unsure whether certain students were present on a particular day: *“I don’t know whether these guys were here on that day”* (T1). A recurring question focused on **how the system detected and defined teacher visits to students**. One teacher asked whether the system recognized her standing between two groups of students, *“If I stand in the middle... I can look at these two kids and these two kids... Would it only assign it to one person? Would it count as a visit for everybody?”* (T2). The same teacher speculated that her short visit durations might explain why certain visits were not registered: *“They don’t start reporting until after 21 seconds... I try to keep them short so that I don’t spend too much time holding their hand”* (T2). Teachers also sought clarification about **how specific analytics metrics were defined**, particularly for nuanced behaviors such as “misuse of time”. One teacher asked, *“What exactly defines a misuse of time?... Is it guessing actions or clicking hints rapidly, next, next, next... clicking to bottom out hints and just type it in?”* (T2).

These instances illustrate how Reflecto not only engaged teachers in interpreting data and raising questions during the process, but also prompted a level of **deeper instructional reflection** and hypothesis forming behind classroom situations and interactions that is not easily achieved through traditional reflection practices without this unique MMLA data display.

10.2.5 (RQ 2.2) How does reflection bring new knowledge to teachers (Surprise)

This section describes the instances in which teachers expressed surprise, defined as *moments when Reflecto’s information either surprised them or contradicted their prior knowledge* (e.g., “The only thing that stands out for me is this [pointing at ‘combine like terms’, ‘make

constant', and 'make variable']..."). Based on the affinity diagram results, quotes coded under "Surprise" were categorized into three overarching themes: 1) **Teachers were surprised by students performing below expectations**, based on the misuse time, accuracy, and the number of problems solved (N = 15). 2) **Teachers were surprised by low-performing students exceeding expectations**, such as demonstrating higher accuracy, lower misuse time, and more persistence than anticipated (N = 19). 3) **Teachers were surprised by their own visits or teaching behaviors** as displayed in Reflecto (N = 2).

Surprised by Students Performing Below Expectations. Teachers were surprised when data revealed that some students performed worse than they had expected. For instance, T1 noted, *"And then... I know this is going to be very shocking for Ms. Jones, but Amy Modins completed no problems."* Similarly, T2 was surprised by a student she considered diligent: *"Dina had a lot of misused time. What surprises me because she works, she's a hard worker, she's focused. So that was interesting. I felt like that was a high number... So I'm really surprised by her"* (T2). T2 often elaborated on these surprises with thoughtful reasoning, such as possible contributing factors: *"Emily, that surprises me. For her, she's pretty on point normally. It's really high for misuse time and low for accuracy. She's normally like an AB student. So that surprises me. But depending on the day... what's going on outside of math class..."* (T2).

Surprised by Low-Performing Students Exceeding Expectations. On the other hand, teachers were sometimes pleasantly surprised when students who typically underperformed actually did better than expected. This included higher accuracy, reduced misuse time, longer working time, more problems solved, and stronger performance from ESL students. For example, T2 remarked: *"Jenny has 0% [misuse time], which is wonderful... Which surprises me because if she gets something wrong, she normally shuts down."* These moments often prompted deeper reflection that may not have been possible without Reflecto's data about students' performance (especially accuracy) and meta-learning state (i.e., misuse time): *"Queenie surprises me in a good way. I mean, her accuracy isn't super high, but her misuse of time tells me she at least attempted to solve them on her own before just clicking through"* (T2). T1 was also surprised when students who complained throughout the session actually completed their work: *"Dreon did well, which was surprising because he complained the entire time"*. Such cases provide a valuable source of reflection that may challenge teachers' prior assumption of a student's ability and traits, *"He was actually killing it. He's not a dummy"* (T1). These surprising moments may equip new, refreshing knowledge about their students to teachers, and in a way encourage teachers to have a growth mindset about students, and promote their professional development.

Surprised by Visits and Teaching Behaviors. Teachers were occasionally surprised by the analytics on their own visits and movement patterns. For example, T1 was amused by the seat map visualization: *“It was like I was in a marathon.”* In another case, T2 was surprised by the low number of recorded visits to a particular student: *“So I’m surprised he doesn’t have more visits.”* We explained that Reflecto initially did not count brief nudges (under 21 seconds) as visits, due to system detection thresholds. In later iterations, we revised the algorithm to lower the threshold to 10 seconds, and this update was explored during T2’s final collaborative reflection session.

In summary, the “Surprise” code captured moments where Reflecto revealed discrepancies between teacher expectations and student behaviors (primarily) or teacher behavior (occasionally). These insights often prompted teachers to reconsider their prior assumptions about students and themselves, highlighting the value of data-driven reflection.

10.3 Results for Research Question 3

In research question 3, we asked: How does the use of an MMLA-based reflection tool impact teachers’ actual teaching practices?

- RQ3.1: What actions or intentions do teachers set following reflection—what do they aim to keep or change?
- RQ3.2: To what extent do teachers carry out the changes they intended—how do their stated intentions compare with their observed classroom behaviors?

10.3.1 What actions or intentions do teachers set following reflection (RQ 3.1)

After reflection with Reflecto, I asked T1 and T2 whether there are any behaviors they’d like to change to improve their teaching practice. T1 usually replied that there wasn’t much change she’d specifically want to make, and that she already knew these students for a long time. T2, despite also being confident in her knowledge of students, specified several visit behaviors she’d like to change.

Though T1 did not articulate specific behaviors to change, she did mention some pedagogical behaviors that she wanted to enforce during the reflection sessions, without specifying whether they were already her practice or her future intentions. These tended to be **broad and strategy-focused**, such as prioritizing visits to struggling students early in lessons, prompting students to write down their thinking before seeking help, and modeling tool use to minimize confusion. These align with general strategies for promoting metacognition, autonomy, and support among learners. In contrast, **T2’s intentions** were more **student- and data-driven**, identifying specific learners and precise adjustments — such as extending

check-ins, increasing visit frequency after noticing low interaction logs, or providing early support for struggling students (Table 10.3.1 for details).

Table 10.3.1 Teachers' Post-Reflection Intentions and Rationales

T-Session	Intention	Supporting Quote	Rationale
T1-1	Prioritize visiting struggling students early	<i>"Head to the strugglers at first."</i>	These students often don't begin independently and benefit from early teacher prompting.
T1-1	Require students to write their problem-solving process before seeking help	<i>"If it's not written on paper ... I'm not coming to you."</i>	Promotes students' accountability and enables the teacher to assess students' understanding.
T1-1	Provide a demonstration on using the tool/system	<i>"I feel like my kids will need a demo ... eliminate those errors and hiccups."</i>	A demo reduces confusion and increases student readiness.
T1-2	Encourage peer-first help-seeking	<i>"Ask peer before me; ask two people before asking me."</i>	This strategy builds student independence and reduces teacher overload.
T2-1	Pay more attention to Sadie (low accuracy)	<i>"...focus a bit more to make sure she understands ... asking the questions."</i>	Sadie's literacy impacts her math; she needs support.
T2-1	Increase visits to overlooked students (Mia, Alex)	<i>"Mia has zero visits ... I'd want to pay attention to her."</i>	Ensures no student is overlooked
T2-1	Extend check-in duration for easily distracted students (Sean, Caleb)	<i>"Sean, Caleb... longer check-ins ... instead of a short nudge."</i>	Longer interactions support deeper focus and prevent distraction-related issues.
T2-1	Spend more time with Bryce (question-asker) early in tasks	<i>"Bryce... question asker... spent more time ... accuracy is great."</i>	Front-loading support scaffolds Bryce's learning effectively.

10.3.2 To what extent do teachers carry out the changes they intended (RQ3.2)

We categorized teachers' intentions into several types. First, teachers discussed some instructional intentions, such as asking the researchers to demonstrate a tutoring software practice modeling problem-solving steps before students start to solve problems, and encouraging students to ask peers before asking the teacher. T2 mentioned time allocation intentions, like adjusting the amount of time spent with certain students.

For T1's intentions, they were more challenging to verify (at least based on the current stage of analysis). As T1 described more general behaviors, and there were limited links from the data to the intentions, validation was often inconclusive. To evaluate T2's intentions and whether her behavior changed accordingly, we compared her intentions with classroom log data across the four instructional days (shown in Table 10.3.2).

For T2's intentions, we extracted the teacher analytics (visit frequency and duration) and students' analytics accuracy, for all the study days and for each individual student that the teacher mentioned (Table 10.3.2). We compared the post-reflection visit analytics (Day 3 & 4) with the pre-reflection visit analytics (Day 1 & 2).

For four students, T2 showed **an alignment between her intentions** and the following behaviors. Specifically, T2 intended to "pay more attention" to Mia, and her visit frequency increased from two on Days 1 & 2 to six on Days 3 & 4. The total visit duration is also longer (from 55 seconds to 166 seconds) (Table 10.3.2). The log data shows evidence of this clear behavioral change. T2 commented on a decline in the student's accuracy over time and linked it to the more challenging material: *"she had cruised through the portions... they were most familiar with and now [was] on sort of newer material"*(T2). Similarly, for Sean and Caleb, T2 intended to have "longer check-ins" with them after checking in the Reflecto data. Both visit durations and frequencies increased post-reflection (Table 10.3.2), suggesting that T2 did adjust her teaching behaviors to align with her intentions. She also noted improvements in both metrics herself. T2 combined the interpretation with the student's learning analytics: *"His misuse of time was 66 [on Monday]... Tuesday... It was 36. So that went down, which is good."*

However, T2 did not achieve her intention for two of the students, and T2 offered related discussion, mostly drawing connections to content familiarity: T2 explicitly intended to "focus on Sadie a bit more" since Sadie had low accuracy. However, T2 did actually not visit her on Day 3 and made only two short visits on Day 4, totaling 46 seconds. These visits were shorter compared to 3 minutes 20 seconds on Day 1 and Day 2 (Table 10.3.2). This suggested that T2 did not follow her intention set for this particular student. T2 acknowledged this student's low accuracy on Day 4, and she attributed it to the unfamiliar content students encountered on Day 4, saying that *"It could be that she got to that newer material and it was something that she definitely was less familiar with"*. In the case of Alex, although T2 had a strong intention to "check in more often", the log data showed the opposite trend, as the visit count and duration decreased over the four instructional days (Table 10.3.2). T2 noted that content familiarity likely influenced student performance: *"Once we got to Tuesday, it might be the content again that they weren't as familiar with."* Taken together, these findings indicate that while T2 implemented her intentions with some students, but not consistently in all cases.

Table 10.3.2 Tracking Teacher Intentions and Follow-Through Using MMLA Data (T2): Visit Counts, Durations, and Student Accuracy Over Four Days for Mentioned Students

Student (Alias)	Teacher's Intention	Visit Count (Day)				Visit Duration (Day) min:seconds				Accuracy (%) (Day)			
		1	2	3	4	1	2	3	4	1	2	3	4
Sadie	Focus on her a bit more	2	6	0	2	1:18	2:02	0:00	0:46	30	47	72	21
Mia	Pay more attention	0	2	2	4	0:00	0:55	0:43	2:14	71	44	66	55
Alex	Check in more often	6	3	2	1	1:46	0:35	0:54	0:30	34	64	71	56
Sean	Want to have longer check-ins	0	–	3	4	0:00	–	1:15	2:14	30	–	48	59
Caleb	Want to have longer check-ins	2	2	–	4	0:34	1:12	–	2:03	61	82	–	60
Bryce	Spend more time with him	3	2	5	4	1:10	0:28	2:51	1:23	59	66	62	83

10.4 Study Results for Research Question 4

In Research Question 4, we asked: What are teachers' perceptions of the MMLA-based reflection tool, and what design feedback do they offer?

- RQ4.1: What features do teachers find most helpful in supporting their reflection practices?
- RQ4.2: What are their views on using the tool regularly over an extended period?
- RQ4.3: What aspects of the tool did not work well?
- RQ4.4: How can the tool be improved to better support teacher reflection?

For data analysis, we analyzed 1) comparison between teachers' actual reflection process with the expected model, 2) teachers' interview and exit survey on feature usefulness qualitatively through affinity diagramming, here, we present the analysis results.

10.4.1 RQ4.1: What features do teachers find most helpful in supporting their reflection practices?

Teacher action reflection that aligns with the expected model. Comparing if teachers' reflection actions align with the expected model (Section 9.1), several aspects were found to work well: 1) After receiving guidance and data-analysis suggestions, teachers were generally able to understand and carry out trend selection, rating, and data interpretation. 2) Teachers

were generally successful in examining the Class Table and Seat Map, sorting and selecting relevant metrics, and engaging in meaningful reflection based on both teacher and student analytics—often connecting these insights to their prior knowledge and memories of students. 3) Teachers effectively validated their existing instructional pedagogy, found support in the analytics, and gained deeper, data-driven insights into their student visit patterns. 4) During collaborative reflection, teachers shared their experiences, supported and built on each other’s insights, and helped one another navigate challenges in data interpretation or tool usage.

To further understand which features best supported teachers’ reflection practices, we analyzed those with the highest usefulness ratings (4.5–5.0) from two teachers, T1 and T2. The results can be found in see Fig 10.4.1.

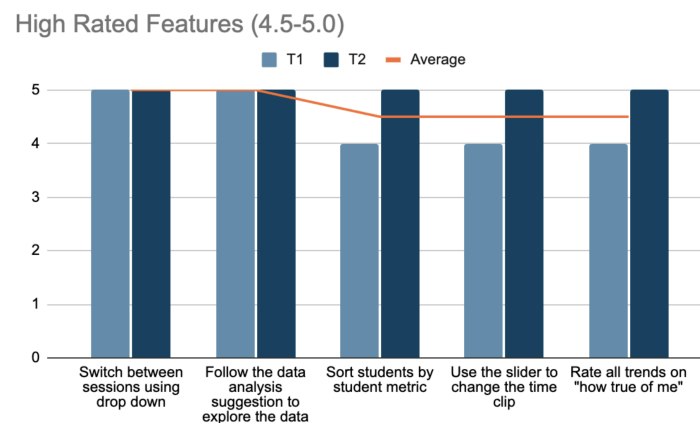


Figure 10.4.1. Highly Rated Features in Reflecto and Teachers’ Respective Ratings

Both teachers rated the feature that allowed them to switch between different class sessions in Class Panel (screenshot illustration in Fig. 10.4.2) to be extremely useful (T1: 5, T2: 5; Average = 5.0). T1 mentioned she thinks this is very useful as she “*would like to see the data on all of her classes,*” while T2 agreed that it “*was super important*”.

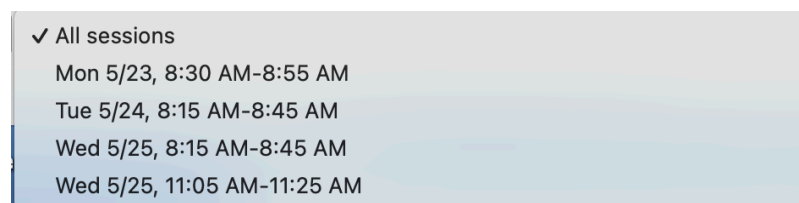


Figure 10.4.2. Feature in Reflecto Allowing Teachers to Switch Between Sessions Using the Drop-Down Menu

Both teachers also rated the feature that allowed them to follow the data analysis suggestion to explore the data (one example screenshot illustration in Fig. 10.4.3) to be extremely useful.

(T1: 5, T2: 5; Average = 5.0). T1 found it “*helpful to get to know the software,*” while T2 found it easy to use the feature in particular within the seat map, claiming: “*It's super easy in particular with the seat map. It's nice to have the direction sort of spelled out for the class table as well. I think it's easier to figure out how to analyze that information in the class table than in the seat map*”.

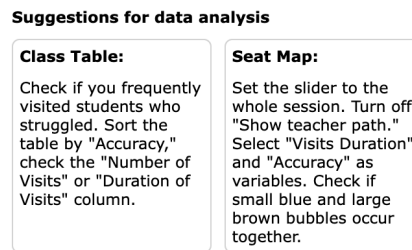


Fig. 10.4.3 Feature in Reflecto Allowing Teachers to Follow Data Analysis Suggestions to Explore the Data

Teachers found the feature to sort students by metrics in the Class Panel (screenshot illustration in Fig. 10.4.4) to be very useful (T1: 4, T2: 5; Average = 4.5). T1 “*like the ability to sort the data*”. For data accuracy, T2 raised concerns about the accuracy of "Idle Time", as she noticed it always showed as 0. Despite her concern on the accuracy of certain data columns (that are continuous zero), she thought the feature is “*super important*” because “*There are some kids that will always look busy when they’re not. And then there are the kids I knew that were just clicking through that were capable*” (T2). Similarly, T1 rated this feature as moderately useful to infer student engagement.

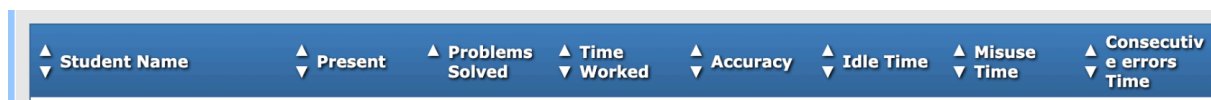


Fig. 10.4.4 Reflecto Feature Allowing Teachers to Sort Students by Metrics

Besides, teachers also rated the feature to use the slider to change the time clip in the Seat Map panel (screenshot illustration in Fig. 10.4.5) to be very useful (T1: 4, T2: 5; Average = 4.5). As T1 mentioned, “*I liked how you could be specific on the time if you needed to be,*” T2 also found it useful for isolating different time frames during lessons, mentioning she would “*be interested to see what that looks like during direct instruction*”.

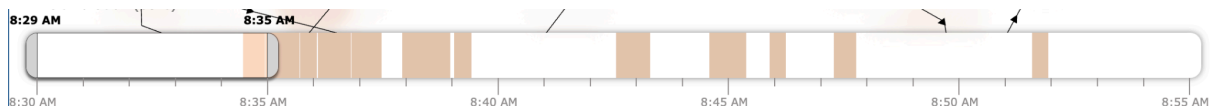


Fig. 10.4.5 Reflecto Feature Allowing Teachers to Adjust the Time Clip With a Slider

Lastly, teachers found the ability to rate all trends on “how true of me” to be very useful. (T1: 4, T2: 5; Average = 4.5). T2 liked that the feature let her compare her initial thoughts about students with what the data actually showed: *“It was nice to be able to see what I thought in the beginning as opposed to what the data actually proved”* (T2).



Fig. 10.4.6 Reflecto Feature Allowing Teachers to Rate All Trends on “How True of Me”

Below, we report additional Reflecto features that received an average usefulness rating of 4.0 out of 5.0 (Fig. 10.4.7). These features were considered very useful for supporting teacher reflection, although teachers’ ratings showed some variation.

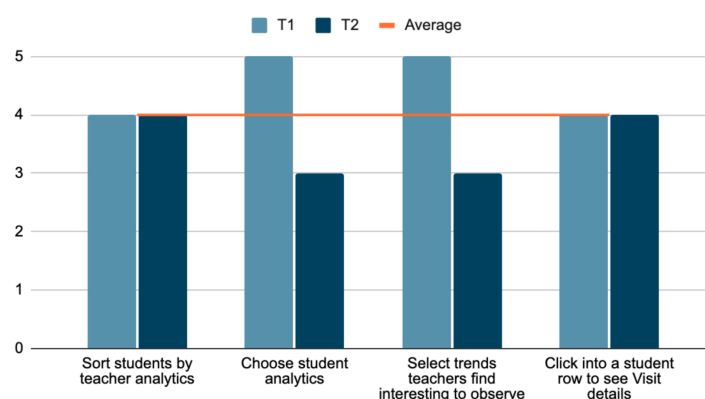


Fig. 10.4.7 Reflecto Features With an Average Teacher Rating of 4.0 for Usefulness

Both teachers rated the features that allowed them to sort students by teacher analytics in the Class Panel (Reflecto screenshot in Fig. 10.4.8) to be very useful (T1: 4, T2: 4; Average = 4.0). Both of them mentioned *“visit count is more important than visit duration”* (T1). T2 emphasized the value of visit frequency, saying, *“I feel like the visit count is probably more important to me than the visit duration, (as) seeing how often it was super important.”* T2 think the visit count can reflect well on her classroom interaction and visit strategies: *“It was interesting because in the beginning, mine (my visit counts) were so low because I kind of just do the quick visits, like nudge them in the right direction, allow them to attempt to work on it. Especially with content that I know that they should be able to work on.”* T2 believes that visit duration still holds value, particularly in shedding light on why certain visits last longer

than others. *“Like with Olivia, I could see that her visit duration was long in the beginning and that was probably a result of getting her started, explaining or reading through directions with her”* (T2).



Fig. 10.4.8 Reflecto Feature Allowing Teachers to Sort Students by Teacher Analytics

In the seat map panel, teachers found the feature that allowed them to choose student analytics through the radio button (Reflecto screenshot in Fig. 10.4.9) to be very useful, though teachers’ ratings differ (T1: 5, T2: 3; Average = 4.0). T1 mentioned that she pays *“Most attention to the ‘Problem solved’ and ‘Accuracy’ because they are the most important metrics”* to her. Relatedly, T2 thought she would *“probably look at things like misuse time and accuracy more.”*; she also noted that the ‘Time worked’ *“was a little bit more difficult to see that information on the seat map.”*



Fig. 10.4.9 Reflecto Feature Allowing Teachers to Choose Student Analytics in the Seat Map Panel

Teachers also rated the ability to select trends they find interesting to observe in the Focus Panel (example screenshot in Fig. 10.4.10) to be very useful (T1: 5, T2: 3; Average = 4.0). T2 found the feature helpful for guiding attention, saying it helped her *“focused on maybe where to keep an eye out to start”* (T2).

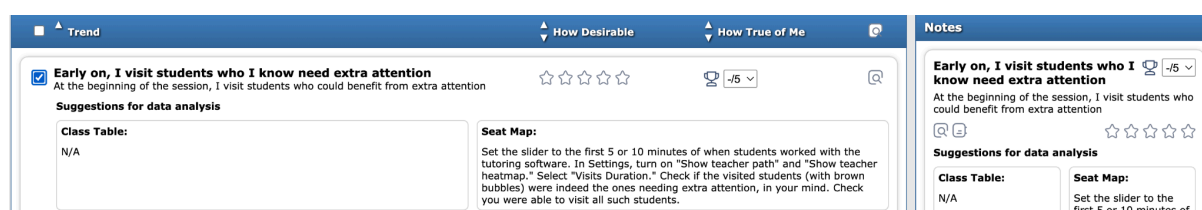


Fig. 10.4.10 Reflecto Feature Allowing Teachers to Select Trends of Interest

Both teachers found the feature to click into a student row to see visit details in the Class Panel (screenshot in Fig. 10.4.11) to be very useful (T1: 4, T2: 4; Average = 4.0). For example, T2 used this feature to support deeper analysis of student progress: *“(This feature) helps in determining how the accuracy adjustments and being able to look at that more. (For example) to see how Amy goes from 40% accuracy to zero after we visit. Now, she is an anomaly all the time, but there were some drop.”* T2 noted that more detailed instructions on this feature may help increase their use of this feature: *“Maybe knowing more about the program would help.”*

yellow b	✓	9	20m	88%	0%	0%	0%	4	1m 55s
yellow r									

Visit Details

orange o bat

Interval Size Before Visit (minutes):

5

Number of Visits

Duration of Visits

Interval Size After Visit (minutes):

5

1

0m 44s

Visit	Duration of Visit	Start Time of Visit	Accuracy 5m Before Visit	Accuracy 5m After Visit	Student Actions 5m Before Visit	Student Actions 5m After Visit
1	0m 44s	8:42:35 AM	40%	43%	15	23

Fig. 10.4.11 Reflecto Feature Allowing Teachers to Click a Student Row to View Visit Details

10.4.2 What aspects of the tool did not work well?

Reflecto Features Based on Ratings. To find out which features teachers found least helpful in supporting their reflection practices, we analyzed the set of lowest-rated features (2.0–3.0 average). The results shown in the following bar chart, Fig. 10.4.12.

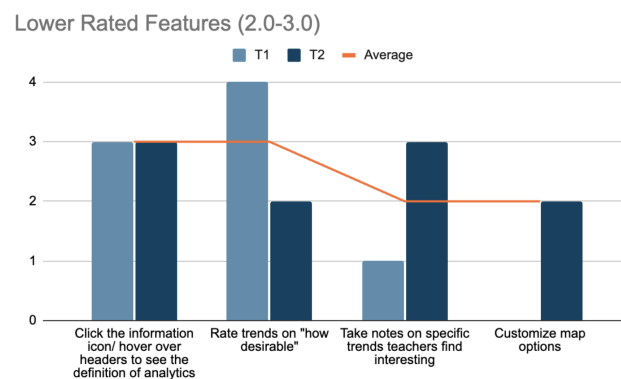


Fig. 10.4.12 Lower-Rated Features in Reflecto and Teachers' Respective Ratings

Both teachers rated the feature that allowed them to click the information icon or hover over headers to see the definition of analytics in the Class Panel (example screenshot in Fig. 10.4.13) to be only moderately useful (T1: 3, T2: 3; Average = 3.0). T1 thought it helped the teachers in the onboarding process, and was *“helpful to figure out how to use the software”*. However, T2 suggested room for improvement, suggesting it would be more useful if it offered *“more information, such as the ability to view more specific details about each title.”*

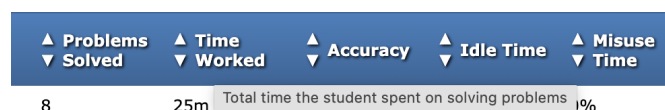


Fig. 10.4.13 Reflecto Feature Allowing Teachers to Click the Information Icon or Hover Over Headers to View Analytics Definitions

Teachers found the feature to rate trends on ‘how desirable’ in the Focus Panel (Fig. 8.4.14)

to be moderately useful. T1 rated the feature very useful, though did not provide detailed reasons. In contrast, T2 said that she *“didn’t find it really useful”*.

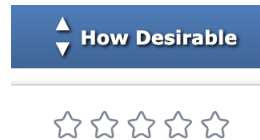


Fig. 10.4.14 Reflecto Feature Allowing Teachers to Rate Trends on “How Desirable” They Are (1–5)

Both teachers gave a low score to the feature that allows them to take notes on specific trends they selected (T1: 1, T2: 3; Average = 2.0). T1 rated this feature not at all useful, saying she was *“not a big note taker”*. T2 reflected that taking notes felt like an extra step rather than part of her natural workflow, and claimed, *“maybe if I were better using the program, it would become more natural to put in notes,”* suggesting that not being fully familiar with how the tool works may have made her less likely to do note-taking.

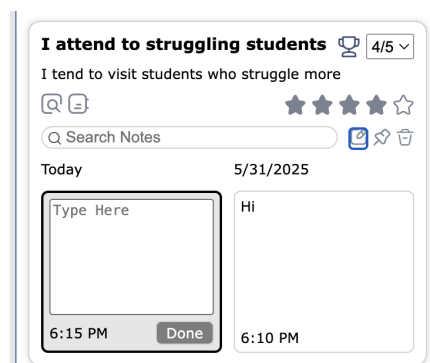


Fig. 10.4.15 Reflecto Note-Taking Feature Allowing Teachers to Record Notes on Selected Trends of Interest

Finally, T2 rated the feature to customize map options in Seat Map to be slightly useful, while T1 did not rate and explained *“didn’t use this because it wasn’t available for me. It is a nice touch!”* (T1: not rated, T2: 2; Average = 2.0). T2 acknowledged exploring the feature but remarked, *“It was a little hard to get a desired output from tinkering within that menu”*. It is worth noting that this panel was initially designed to be only used by the programmer when tinkering with the tool, but we explored with teachers to see whether it is useful. The results indicated that the design could benefit from being made clearer/ user-friendly, if it is kept in future iterations.

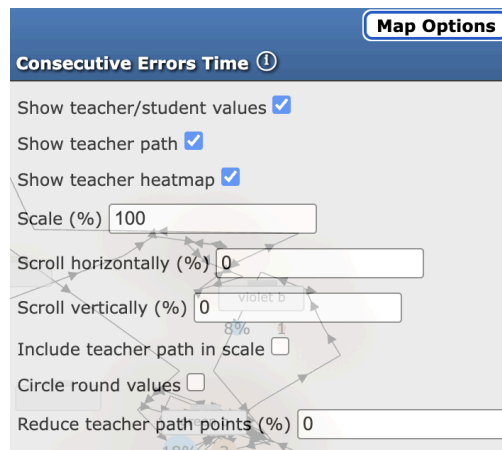


Fig. 10.4.16 Reflecto Customization Feature for Map Options

10.4.3 What are their views on using the tool regularly over an extended period?

Both teachers saw value in using the tool regularly over a longer period of time, with slightly different perspectives. T1 saw long-term potential for improving her teaching and helping her understand students' behavior. She felt that Reflecto helped **reinforce and validated her existing knowledge and intuition**: *"I definitely think (...) it helped reinforce, like, what I kind of already knew, it (Reflecto) could add on as far as reinforcing what I'm already teaching, and then help me gather more data, which I think is always good."* T2 envisioned herself using Reflecto regularly to better **understand her own movement and instruction patterns**: *"I would be interested to see how I circulate during direct instruction because I feel like a lot of times [I] spend just standing up in front of that board."* Besides, she wants to use Reflecto to see how she moves around the classroom and **teaches differently for different content**: *"Seeing how that has changed my instruction or my practice of delivery would be interesting. Seeing how my students, when it's newer content how I circulate the room and how it affects them as opposed to when it's something that they absolutely know how to do (or doing things on their own)".* Given that time was a key consideration for reflection, T1 pointed out rather than a one-time thing, having **recurring reflection every nine weeks** could be beneficial to see long term changes: *"Instead of just doing it twice, maybe do it every nine weeks to see what I can change."* Additionally, related to broader adoption at her school, T2 raised concerns that it would depend on individual teacher attitudes towards edtech: *"Some teachers [are] willing to use it (Reflecto) and other teachers are not. They want to do nothing with technology"* (T2). In sum, teachers are interested in adopting this for a longer time with different purposes and preferred frequency.

10.4.4 How can the tool be improved to better support teacher reflection?

We analyzed interviews, surveys, and teachers' actions in reflection to answer this research question:

Comparing teacher action reflection with the expected model (in Section 10.1.1), several areas for improvement were identified:

1. **Trend Selection:** Teachers in our study tended to explore mostly highly desirable trends. While this is not inherently good or bad, a consistent focus on only positive patterns may limit the depth of reflection and increase the risk of confirmation bias, especially for experienced teachers. Encouraging teachers to also examine undesirable trends could help surface behaviors that require attention or change.
2. Teachers were sometimes limited in their ability to recognize counterexamples, which take time and require familiarity with both the tool and the data. For example, we observed that T2 demonstrated a more data-driven approach during reflection. One contributing factor may be her extended exposure to the tool—she gained familiarity with Reflecto from collaborative reflection with T1 before reflecting herself, and she was involved in two additional individual data exploration sessions. We found that when we identified and presented T2 directly with counterexamples, she offered possible interpretations that may challenge her previous assumptions about the classroom.
3. To promote rigorous reflection to uncover new instructional goals and opportunities for classroom improvement, it may be beneficial to ensure a more comprehensive examination of every student's data in each class. This helps guard against overlooking quieter or less memorable students. This approach exemplifies human–AI synergy: while humans tend to remember the most salient, extreme, or vocal cases, machines can complement this by systematically examining all data points in a fair and unbiased manner.
4. The note-taking function (Fig. 10.4.15) was used only sparingly during the study and was rarely revisited or rediscussed. This suggests an opportunity to rethink how this feature could better support teachers in logging data-driven insights that can be reused in future reflection sessions, and in having note-taking seamlessly integrated into their reflection workflow.
5. Finally, there is room to deepen the level of discussion to go beyond describing what happened and engage more in analyzing instructional strategies and co-developing solutions to common challenges. That said, the extent of this collaboration may depend on how aligned the teachers are in terms of teaching style and subject area.

We also complemented results for this question with evidence from the interviews and surveys:

Seat Map. Both T1 and T2 thought the current zooming and navigation experience in the Seat Map made it harder to explore the data effectively (Fig. 10.4.17). T2 suggested that instead of adjusting zoom with percentage controls, it would be more natural to use finger gestures or mouse dragging: *“I would say being able to scroll using like your fingers to slide it over or something or your mouse like to move your map might be easier than manipulating it in there with percentages.”* T1 imagined being able to click on a student or a portion of the map to expand it: *“If we clicked it all, and then this would become the full screen... expand the window. And then with the zoom-in function, I think it would be way better”* (T1). T2 appreciated the design of the Seat Map for visualizing her movement but found it difficult to interpret data: *“The seat map, I liked looking at it to see my path, to see how I was. I just felt it was more difficult to use the bubbles than it was in the class table to get that information... there's so much in terms of my footpath that it's hard to see the bubbles and who goes with what bubble goes with who.”*

Class Table. Both teachers expressed a preference for using the Class Table over the Seat Map when it came to interpreting data. For instance, T1 found it especially useful because the numbers were easier to read and the layout was not visually overwhelming: *“As far as Reflecto, I thought the most useful was the class table, because it actually showed the numbers and everything wasn't all bunched up... So I like the class table.”* T2 also expressed the same preference, mentioning that the Class Table helped her interpret data more efficiently than the visual representation using bubbles in the Seat Map.

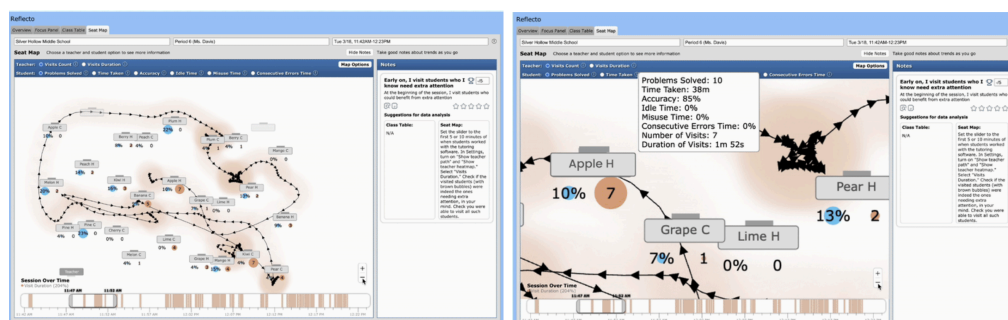


Fig. 10.4.17 Current Zoom Function in Reflecto

Focus Panel. T1 expressed interest in making the Focus Panel more flexible by allowing teachers to track their own personalized teaching trends—not just select from the system-provided ones. She imagined teachers could add their own trends that mattered to them and reflect on them consistently over time: *“It might be nice to have... just have a trend that you could add on to there and have the flexibility to, like, track that.”* (T1)

10.5 Study Results Summary and Discussion

10.5.1 RQ1: How Teachers Use Reflecto for Individual and Collaborative Reflection

Results Summary: Trends Reflected and Desirability (RQ1.1). Teachers consistently selected **high-desirability trends**—such as early support for struggling students, strategic short visits, and broad spatial movement. These trends appeared across multiple sessions, indicating a stable focus on verifying instructional practices. **Counterexamples**—students whose behavior or outcomes didn’t align with the trend (e.g., a student visited early who still struggled)—were sometimes noticed but typically not explored in depth. However, when prompted, T2 engaged thoughtfully with these mismatches and offered contextual explanations or hypotheses (e.g., misunderstanding directions, individual learning habits, or task-related confusion). This suggests that counterexamples were more likely to be addressed when explicitly surfaced. **For the conclusion drawn for trends reflected (RQ1.2),** teachers frequently rated trends as **highly true of themselves** (4–5 out of 5) when they felt consistent with their classroom approach, often drawing on teaching memory and impressions from the lesson. In exploring Reflecto’s analytics, teachers commonly focused on a few salient students, rather than scanning the full set of student data. Overall, trend ratings reflected teachers’ perceived consistency with the trend and alignment with their teaching approach. In terms of the **data analysis process**, teachers used Reflecto’s Class Table and Seat Map to examine patterns related to their selected trends, filtering by metrics like visit frequency or duration, students’ accuracy, and misuse time. Their analysis often centered on memorable or representative students, using classroom memory as a primary lens for interpreting data. Teachers occasionally used Reflecto’s **note-taking** feature to record concise instructional beliefs and pedagogical values. As for the collaborative **reflection process (RQ1.3),** teachers engaged in **supportive** dialogue that included affirming one another’s insights, sharing classroom stories, and discussing contextual differences such as **grade-level needs** or tool usage. These discussions exposed common teaching challenges and allowed teachers to share rich pedagogical knowledge. One potential way to make reflection more actionable and to encourage teachers to propose suggestions is to have teachers who teach the same grade level and content co-reflect together.

Discussion: *What is surprising about the results & why they matter:* while we may expect teachers to explore trends of different desirability, they were surprisingly unanimous in selecting highly desirable trends. This reveals an interesting insight that experienced teachers (at least in this study) may lean towards examining good pedagogical practices in collaborative reflection. Another surprising insight is that teachers drew conclusions of desirable pedagogical trends being true to them consistently, and these experienced teachers were more likely to affirm, rather than criticize, their own and their colleagues’ teaching

practice. The study findings about these inclinations *matter*, as they reveal challenges in how to avoid “confirmation bias” in reflection.

10.5.2 RQ2: How Reflecto Influenced Teachers’ Understanding of Their Teaching and Students

Results Summary. Qualitative analysis from “Confirm Knowledge” and “Recall Knowledge” concerns *how Reflecto relates to and interacts with teachers’ existing knowledge* (RQ 2.1). On the one hand, Reflecto validates teachers’ existing understanding of student *performance, behaviors, and their own instructional strategies*. Metrics like accuracy, misuse time, and task completion informed and confirmed teachers’ prior knowledge about students’ performance and learning behaviors (e.g., help-seeking tendencies). Reflecto also reinforced teachers’ awareness of teaching strategies, including classroom movement, short visits, and proximity-based support. These moments of confirmation, empowered by Reflecto, reinforced teachers’ prior beliefs about their pedagogical routines and knowledge of students. In these reflections, most frequently, teachers engaged in recall of classroom events, student traits, and instructional practices. These reflections were usually grounded in memory but triggered by Reflecto data. Teachers recalled student attendance, personality traits, learning styles, and performance, as well as their own teaching behaviors, such as visiting principles or what happened during a specific visit. Teachers also recalled broader classroom dynamics including content difficulty, discipline issues, or environmental constraints.

On the other hand, do these reflection sessions bring new knowledge to teachers (RQ 2.2)? Affinity diagramming for quotes under “Learn Info”, “Surprise” and “Question” showed the answer is yes. Reflecto facilitated data-informed observations about students and teaching not previously recognized. Utterances under “Learn Info” show teachers used Reflecto to uncover new insights by noting student metrics such as accuracy or number of visits, identifying individual or class-wide patterns across different sessions (e.g., change, improve, decline), as well as gaining more awareness of their own visit patterns and durations. For quotes under “**Questions**”, teachers raised questions about the accuracy, definition, or algorithm design of Reflecto’s analytics. One of the most valuable sources of reflection in my eyes is their *data-driven hypothesis for reasons behind student analytics and behavior*. Teachers hypothesized in depth about possible reasons, including content and curriculum, language barriers, students’ emotional and physical states etc. Teachers also reflected on their own instructional and visit decisions, such as why certain students received more attention. These questions and hypotheses highlight the tool’s role in sparking critical reflection beyond surface-level observations. Reflecto also prompted genuine **surprises** in teachers, through revealing cases of underperformance from typically strong students, as well as struggling students who exceeded expectations in accuracy, persistence, or task completion. Reflecto occasionally revealed unexpected insights about teachers’ own behaviors, such as lower-than-expected visit counts or concentrated movement in areas. These surprises

represent valuable insights that would not be uncovered without this MMLA-based tool.

Discussion: *What is surprising about the results & why they matter:* It is surprising to find that the “Recall Knowledge” takes up the majority of the quotes coded in the reflection sessions transcripts. These recall of past memories and events are in nature more descriptive rather than evaluative. Another surprising result is that students’ *differentiated help-seeking* strategies were frequently discussed across different categories. These indicated that students’ own learning traits and whether they seek teachers’ help play an important role when teachers make decisions about distributing their time in visiting different students.

10.5.3 RQ3: How Does Reflecto Influence Actual Teaching Practices?

Results Summary. From video transcript analysis of the reflection sessions, we analyzed the intentions to change practice that teachers set following reflection (RQ3.1). T1, albeit not specifying behaviors she’d like to change, emphasized instructional strategies aimed at improving student independence and accountability. T2, who participated in additional individual data exploration sessions, set intentions that were more data-driven and student-specific. Grounded in Reflecto’s analytics, she identified individual students and described tailored instructional adjustments such as increasing visit frequency, extending check-in duration, or offering support earlier in the task. T1’s reflections tended to be broad and more challenging to verify (partly limited by the analysis timeline and data accessibility). To answer the extent to which teachers carry out the data-driven changes they intended to make (RQ3.2), we extracted the T2’s teacher analytics (visit frequency and duration) and students’ analytics accuracy, for all the study days and for each individual student that the teacher mentioned (Table 10.3.2). Findings show T2’s intentions were followed through for four out of six students who she intended to change behaviors about (details in 10.3). For example, one successfully carried-out intention was that, after T2 noticed that Sadie had received little support, T2 said, “*Sadie, she’d be somebody I’d want to pay attention to.*” On Day 4, Sadie received the highest number of visits (4), indicating a meaningful shift. When T2 was prompted to revisit original intentions, she drew on factors such as content difficulty change to interpret learning analytics, “*Only at 21 [percent accuracy]... it may be that she got to that newer material she was less familiar with*”. Results for RQ3 showed that an MMLA-driven reflection tool can support teachers to set data-driven intentions. When intentions are framed around specific students, they hold great potential to **increase teachers’ awareness** and lead to observable instructional shifts.

Discussion: *What is surprising about the results & why they matter:* The most surprising result from this research question is that the two teachers set quite different intentions. The specific, data-driven intentions set by T2, in my view, signify that reflection sessions can bring about actionable benefits and potentially longer-lasting impacts (e.g., by challenging

how she wanted to interact with certain students). A logical hypothesis is that these more specific intentions may partly stem from the longer data exploration time T2 had during the individual reflection sessions, which afforded her greater immersion and deeper familiarity with the data. This raises an important direction for future research: exploring, through rigorous experimentation, whether additional individual reflection sessions lead teachers to set more data-driven and student-specific intentions.

10.5.4 RQ4: Teachers' Perceptions and Design Feedback on Reflecto

Results Summary. This research question examined how teachers perceived the MMLA-based reflection tool based on analysis of interviews, surveys, and comparison between expected models and the teachers' actions. Findings are organized around feature usefulness (RQ4.1), long-term use (RQ4.2), limitations (RQ4.3), and design suggestions (RQ4.4). For **useful features in Reflecto (RQ4.1)**, the most valued features were those that scaffolded exploration and comparison, including the ability to switch between sessions, follow data analysis suggestions, sort student metrics, and adjust time windows. T2 appreciated its potential to challenge her assumptions, stating: *"It's very valuable because sometimes you think you're doing one thing, and then you see the data and you're like, 'Oh, maybe I'm not doing that as well as I thought.'"* This suggests that the tool supported both confirmation and recalibration of instructional beliefs. Both teachers *preferred the class table* for tracking metrics like visit counts, duration, and accuracy. For **long-term use of the tool (RQ4.2)**, T1 viewed Reflecto as reinforcing her knowledge of her classroom practices, and suggested it could be used every nine weeks to track changes. T2's long-term interest in using Reflecto centered on analyzing her movement during direct instruction or when introducing new content. For **tool limitation (RQ4.3) and room for improvement (RQ4.4)**, teachers rated the note-taking feature and map customization as least useful, primarily due to a lack of habit (T1: *"Not a big note taker"*) or usability issues (T2 found the map menu hard to navigate). Both teachers commented that while the seat map was visually helpful, it was harder to extract detailed insights—especially when classroom layouts were dense or screen space was limited. This suggested that spatial data representations must balance visual intuitiveness with clarity and data legibility. For design recommendations, T1 desired allowing users to *define their own instructional trends* to track over time, supporting **personalized reflection goals**, while T2 requested clearer controls for navigating the seat map. These suggestions emphasize that teachers may seek more customizable reflection capabilities when using MMLA tools.

Discussion: *What is surprising about the results & why they matter:* One interesting finding is that teachers expressed interest in adding personalized trends to track their own behaviors, which points a good direction towards how MMLA-based reflection tools can be made more teacher-centric and personalized. It is also encouraging to learn that both teachers think such

MMLA-based reflection tools can be used for a long time. The results shed light in unique ways they may want to leverage the reflection tool, e.g., in a customized frequency to understand instructional practice of particular interest to them (e.g., during direct instruction & covering new content).

10.5.5 Study Discussion, Contribution, and Limitation

A teacher's positioning in the classroom plays a crucial role in managing the environment and engaging with students (Pozdniakov et al., 2022). Collaborative reflection may offer advantages over individual reflection, through cognitive conflict or the support of capable peers (Kim & Lee, 2002). While past work has explored video-based reflection (Romano & Schwartz, 2005), prototype tools (Prieto et al., 2020), and student-facing dashboards (Yan et al., 2024), this study demonstrates how a teacher-facing tool can meaningfully integrate location data about teachers' movement with student learning data from tutoring software to support reflection. This combination enabled teachers to examine what students were doing (through accuracy, misuse, and other learning analytics), and how their own physical presence related to engagement and performance (e.g., whether visits improve students' accuracy). This study builds on prior research into technology-supported teacher reflection and contributes to both theory and design in the context of multimodal learning analytics (MMLA) and K–12 math instruction.

There are close connections between our tool and the literature guideline. For example, the design of Reflecto aligns with the design guideline by Prieto et al., for designing the socio-technical pair (designing the technology and routine) together, to help automate reflection practices within teachers' routines. Specifically, Reflecto supported a reflection routine of: trend selection → data analysis to examine whether the trend is followed or avoided (making data-driven observations and taking notes) → rate and reflect based on the trend. In the same guideline, Prieto et al. argued for “**designing for overload**” to minimize the risk of forgetting data collection and reflection moments (Prieto et al., 2020). Our tool design also reflects this student-centric value and effort to automate data collection, as we integrate the visit detection algorithm, pipeline, making the collection of student learning and metacognitive data seamlessly into the reflection tool. Their guideline also suggests teacher reflection should require active participation from students (who should be centered as reflection is ultimately meant to impact them) (Prieto et al., 2020), which is also aligned with our tool design, which treats students' learning analytics as an integral part of multimodal data analytics. **Our work, through a field study in an authentic classroom, demonstrates that an MMLA-based teacher reflection tool (combining ITS log for students' learning performance and behaviors as well as teacher location), holds great potential to spur useful reflection. Reflecto helps teachers validate and challenge teachers' existing knowledge about their teaching strategies and about their students. Moreover, our study**

demonstrates that such a tool holds great potential to lead to observation change in classroom practice of teachers, if appropriate scaffolding were in place.

Teachers valued features that scaffolded exploration (e.g., switching sessions, time filters, sort-by-metric), and desired more customizable trend tracking, easier navigation in the seat map, and embedded explanations. These insights can inform the design of teacher-facing MMLA tools that better support diverse reflection habits and instructional goals. Several of our findings provide concrete evidence that extends existing literature on teacher reflection and the design of relevant technologies, offering both theoretical and design implications as follows:

Teacher-centered, question-driven reflection. Prieto et al. highlighted the importance of paying attention to data ownership, for example, allowing teachers to **personalize the behaviors and items to reflect on** (Prieto et al., 2020). Similarly, we found a participating teacher’s interest in being able to add a personalized trend to track their own behaviors. This also connects to prior literature which advocate for personalized, teacher-driven data collection methods that empower educators to define what data is useful and how it should inform their reflection (Saar et al., 2018). For example, specifically, in our contexts, we can consider having teachers be more involved in the co-design process of the reflection tool to customize the trends and visit detection algorithms (e.g., setting time and space thresholds that they want to set for the system to register their proximity as visits in their particular classroom contexts). In literature, there is also an emphasis on “**question formulation**” as the foundation for sensemaking, decision making, and taking pedagogical actions. For example, prior work has proposed a sensemaking framework for teachers which starts by identifying the educational questions that the analytics can address (Wise & Jung, 2019). Some other work consider the LA process as a teacher-inquiry loop in which a teacher’s questions can drive the customisation of the enabling technologies (Rodríguez-Triana et al., 2018; Sergis & Sampson, 2017). In our work, the Reflecto tool carries an implicit question: *“Do teachers’ actions align with the chosen trend they are interested in exploring?”* However, it may be worth investigating how to make these questions more explicit and the reflection process more question-driven. For example, one possible addition could be a step before the reflection phase, with or without a technological prompt, that asks teachers: *“What questions would you like to answer through the teaching and learning analytics?”* or *“What are some questions you are curious to explore in your class?”* This could potentially make Reflecto and the collaborative reflection process more conducive to setting intentions and more helpful in overcoming confirmation bias, especially for experienced teachers.

Human-AI Synergy in Reflection. Reflection is dependent on the teacher’s ability to capture and analyze high-quality data about classroom practices, students’ experiences, and outcomes (Phillips et al., 2023). Kitto et al. have stressed the importance of embracing

imperfect analytics to stimulate more critical reflection and scaffolding for learners to “start a conversation” (Kitto et al., 2018). Through our study, we found that surprises and questions arose when there were discrepancies between teachers’ existing beliefs and memories and what the analytics showed. This helps us consider how to place Reflecto’s value and address the potential inaccuracies and cognitive dissonance, so as to empower the teachers to use the data to support meaningful reflection and conversation starting points, rather than a prescriptive evaluation of their behaviors. Just as Kitto et al. argued, the machine’s limitation may be compensated for by the human’s intelligence, as the burden is placed on the learner (in our case, teachers) to make sense of the analytics. This is also aligned with Alfredo et al.’s study, which emphasized the value of human-in-the-loop for multi-modal analytics to support reflection (Alfredo et al., 2025). They argued that the AI should complement rather than replace human judgement, to enhance the safety, reliability and trustworthiness, and potentially create more responsible Multi-modal Teaching Analytics (MMTA) dashboards (Alfredo et al., 2025). For example, specifically in our contexts, we can consider future work having teachers join the co-design process of the reflection tool, customize the trends, and visit detection algorithms (e.g., time and space threshold that they want to set for the system to register their proximity as visits in their particular classroom contexts).

Advanced Technological Support for MMLA-based Sensemaking & Reflection. Technological support can be important for the interpretation and sensemaking of learning analytics (Pozdniakov et al., 2022). In our study, we also found that making sense of analytics—especially when multimodal learning analytics (MMLA) become visually complex—is both important and sometimes challenging, yet it is a fundamental prerequisite for effective reflection (Pozdniakov et al., 2022). This finding suggested that future designs could scaffold the sensemaking process by enhancing dashboard visualizations with visual elements that emphasize only the relevant data point, accompanied by explanatory text (Pozdniakov et al., 2022). In an analytic-driven reflection tool, for example, through using metaphoric visualizations may aid in how students experience the content (Molenaar et al., 2020), e.g., in our case, showing an idling icon to alert teachers who have not been engaged. Furthermore, Phillips et al. argued that NLP, LSTM, natural language processing (NLP) and deep learning (specifically LSTM networks) can support in-service teachers’ ongoing reflection by automating the collection and analysis of instructional data without requiring heavy manual effort (Phillips et al., 2023). GenAI holds the potential to reduce teachers’ reflection workload in sense-making and overcome confirmation bias.

Limitations & Future work. There are several limitations: 1) This study involved only two middle school math teachers. While the depth of qualitative analysis yielded rich insights, findings may not generalize across teaching populations, subject areas, or institutional contexts. Broader validation is needed with more diverse educators and classroom environments. 2) The study involved iterative improvements to the reflection protocol and

tool configuration in response to teacher feedback. For example, only one teacher (T2) received individualized reflection sessions, had access to example/counterexample prompting, and used a modified visit threshold. While these changes improved engagement, they limit cross-participant comparisons and may have influenced how teachers used the tool. 3) Reflecto was a prototype in active development, and not all features were equally polished or fully functional. For example, the seat map had usability challenges, and the trend rating and note-taking features were not consistently used. These constraints may have limited the full potential of teacher interaction with the tool. 4) Parts of the collected dataset, including Reflecto and ITS logs as well as classroom video recordings, have not yet been comprehensively analyzed, leaving potential insights unexplored. 5) Lastly, the teacher reflection sessions were facilitated by a researcher following a protocol; it remains to be studied how teachers may use the tool independently in their own reflection practices.

10.6 Study Conclusion

This study examined how a multimodal learning analytics (MMLA) tool—Reflecto—can support teacher reflection in real-world middle school math classrooms. Through detailed analysis of teacher behaviors, reflections, and tool interactions, we found Reflecto both confirm and extend educators’ pedagogical knowledge and insights of their students and own teaching. Reflection led to several student-level interaction intentions followed through by teachers, demonstrating the potential for reflection to transform into actionable behavioral change. Together, these insights highlight the potential for MMLA dashboards to act not just as data displays to inform and challenge teachers’ understanding, but as mediators of teachers’ improved practice. While the study’s findings are limited by its small sample size and the tool’s developmental stage, they contribute to a growing understanding of how data can meaningfully inform instructional reflection and practice.

Future work should expand to more diverse teaching contexts, refine features that support more human-in-the-loop, teacher-centered, and question-driven reflection tool design, as well as explore further advanced technology (e.g., GenAI) scaffolding for sense-making and reflection. Future work can explore long-term integration of such tools into teacher professional development cycles to better align MMLA systems with the complex nature of teaching and reflection.

CHAPTER 11 SUMMARY AND THESIS CONTRIBUTION

This thesis provides several novel contributions to the field of AI in education, specifically aimed at enhancing teacher support and classroom dynamics through customized, differentiated, and reflective learning environments.

11.1 Customization of Instructional Materials and Technology

In this thread of work, novel contributions of my thesis mainly include 1) empirical results and experiments with the crowd-pipeline for instructional materials improvement, and 2) analysis of authentic educational log data, and insights on the simulation of the feasibility of pairing policies.

With regard to instructional materials, this thesis examined scalable customization of instructional materials, focusing on mathematical hints in Intelligent Tutoring Systems. The results show that using minimal expert guidance in a crowdsourcing pipeline can significantly reduce the time burden on teachers, while also revealing that experts' preferences for hints vary widely. These *findings* highlight both the potential and the limitations of minimal guidance: although it yields efficiency gains, additional scaffolding is needed to ensure accuracy and relevance in crowd-sourced revisions. **Theoretically**, this work extends crowdsourcing research by shifting the focus from content generation to content revision, offering new insights into expert-crowd dynamics and the enabling conditions for successful crowd pipelines. It contributes to theories of customization in educational technology, the diversity of expert preferences (especially for instruction materials of ITS math hints), and the need to better capture and communicate these preferences in system design. **Practically**, the study demonstrates a pathway for reducing teachers' workload by outsourcing certain revision tasks to the crowd, freeing teachers to focus on more critical pedagogical responsibilities. It provides design guidance on when scaffolding is necessary to maintain the quality of crowdwork, in order to increase teacher satisfaction for instructional materials customization.

With regard to technology (pairing algorithm), this thesis evaluates the feasibility (in the sense of student coverage) of four policies that dynamically pair up students, through a novel simulation process for offline evaluation based on authentic classroom data. The *findings* show that dynamically pairing students based on their in-the-moment wheel-spinning status yields good feasibility for struggling students, even under moderate restrictions on allowed pairings. However, a trade-off emerges: the stricter the required knowledge distance between partners, the less feasible the policy becomes. Importantly, any fixed instantiation of a policy does not fully account for class variability and extreme cases—policies perform differently

across classrooms and sessions. These results suggest that policy feasibility may be optimized by gradually loosening parameters or through classroom-level customization.

Theoretically, this work extends the literature on learning group formation by investigating user-centered, dynamic pairing policies, including the novel consideration of students' in-the-moment wheel-spinning status. It addresses a gap in research on dynamic intra-session group formation by showing how feasibility shifts with classroom context and policy parameters (Amara et al., 2016). Methodologically, it contributes the **SimPairing** approach, which uses authentic classroom data and algorithms as design materials to augment researchers' and designers' intuitions—building on and extending prior work such as Replay Enactments (Holstein, Harpstead, et al., 2020). **Practically**, this work contributes design directions for orchestration tools that manage student transitions between individual and collaborative learning. In addition, the SimPairing approach offers practitioners and developers a scalable method for evaluating pairing policies, with potential applicability to other online educational software that capture transaction data.

Limitations: For the hints generated by the crowd-pipeline, there was no evaluation of their impact on the learning outcome in the classroom. For the pairing algorithm, the analysis focuses exclusively on the feasibility of the dynamic pairing policies, which, while important, represents only one aspect of their value.

11.2 Dynamic Differentiation between Individual and Collaborative Learning.

In dynamic combinations of individual and collaborative learning supported by a human–AI co-orchestration ecosystem, I *have several findings across several studies*: The teacher preference study (surveying 54 math teachers) validated earlier design research and classroom studies (Echeverria et al., 2020; Holstein et al., 2019b; Olsen et al., 2020; Santos, 2012), confirming that teachers want to retain the final say over student pairings, provide backup help sources, front-load preparation tasks, and minimize orchestration load. The prototyping study extended these insights into authentic classrooms, showing that dynamic transitions are both feasible and pedagogically valued—especially by teachers, who saw them as a way to achieve more differentiated learning tailored to pace and learning mode. At the same time, challenges emerged: identifying the right timing, partners, and content for transitions without overloading teachers; ensuring students engage productively during collaboration; and preventing disruptive or unproductive task-switching. The controlled classroom experiment compared a *Dynamic* condition—in which students experienced personalized, teacher-orchestrated transitions that were dynamically differentiated in real time—with a *Standard* condition, in which transitions were pre-arranged and occurred simultaneously for all students. Learning outcome analyses showed that students improved on procedural test items but not on conceptual ones, with no significant differences in overall

learning gains between conditions. However, role-based effects emerged: students in solver roles made greater conceptual gains than those in tutor roles.

Theoretically, these studies contribute to the literature on human–AI co-orchestration for differentiated learning. These studies expand the literature on the feasibility, value, and impact of dynamic differentiation. This work extends theories of human–AI co-orchestration by showing that effective orchestration of dynamic transitions requires balancing teacher agency, student needs, and AI support. Additionally, by investigating stakeholder preferences (teacher survey), feasibility (classroom prototyping activities) and learning impact (controlled classroom study), this series of work provides a lens for how dynamic differentiation can be realized and assessed in classrooms.

Practically, this body of work offers design implications for future orchestration tools, as to how to design AI-based support tools that respect teachers’ boundaries and preferences, minimize teacher input burden and orchestration load, and design for smooth, non-disruptive student transitions between activities. These implications provide actionable guidance for developing classroom-ready orchestration systems that combine human and AI strengths to achieve differentiated learning at scale.

Limitation: The limitation of the work is that the teacher participants are based in the US. Though the human-AI co-orchestration tool is not built for a specific subject, I have only evaluated it in the context of middle school math ITS. These may limit the generalizability of results to other demographics and subject context.

11.3 MMLA-based Teacher Reflection Tool

In this thread of work, in the initial need-finding study, I studied teacher preferences regarding MMLA, and found that teachers expressed the strongest interest in data about their ways of motivating students and interactions with students (teacher analytics) and students’ learning and progress (student analytics). I also found that teachers were generally more open about log data collection and much more accepting of students’ log data than teacher log data. Lastly, I found that teachers are most willing to share data with principals and colleagues who teach the same subject (e.g., math). They are less willing to share their data than their students’ data.

Following the investigation about teachers’ MMLA preferences in reflection, I conducted testing, prototyping and eventually an exploratory classroom study for Reflecto with authentic teachers and students. Reflecto, is a multimodal learning analytics (MMLA) tool (built by colleagues) that supports teacher reflection. Study findings include: teachers were surprisingly unanimous in selecting highly desirable pedagogical trends, and drawing conclusion that the trends were aligned with their practices, suggesting an inclination to

affirm rather than criticize their own classroom practices. In reflection, teachers most frequently described classroom events and student behaviors (including help-seeking behaviors) recalled from memory. Teachers expressed interest in **personalizing the reflection tool**, such as adding their own trends to monitor and using it at customized frequencies over the long term.

Theoretically, our work extended the literature of MMLA use for teacher reflection, by not only uncovering teachers' interest in MMLA data collection and sharing, but also demonstrating the potential of MMLA-based reflection tools to impact teachers' knowledge by validating their practice, challenging their prior assumptions about students, and leading to meaningful interaction change in the classroom.

Practically, our study contributed actionable design insights and implications for the design of MMLA-based reflection tools. Such implications include, for example, to counter potential confirmation bias, tools could incorporate prompts or features that nudge teachers toward examining less desirable trends or unintended practices. Another implication is, arranging longer or repeated reflection sessions may be helpful for enabling teachers to generate more specific and actionable intentions.

Limitation: This study has several limitations: it involved only two middle school math teachers, limiting generalizability; the reflection protocol and tool configuration were iteratively adapted (e.g., individualized sessions, prompting, visit thresholds), reducing cross-participant comparability; and Reflecto itself was a prototype with unfinished features that may have constrained teacher interaction. While these factors provide valuable insights into feasibility and engagement, they also restrict the extent to which findings can be generalized or interpreted as definitive evidence of effectiveness.

11.4 Conclusion

In conclusion, leveraging various methodologies, including human-centered design and research, data simulation and analysis, as well as empirical field studies or controlled experiments, this thesis makes contributions to making the AI-based learning environments (specifically, teacher-support tools in context of ITSs) more customized, differentiated, and reflective.

The broad areas that this thesis research concerns are crowdsourcing (which is considered to be part of social computing), learning sciences and technology, and human-AI interaction & collaboration. This thesis research has a strong focus on designing and understanding from the teachers' perspective or building tools that could serve as supporting teachers. The ultimate goal of a learning environment is to demonstrate improved learning outcomes, making students' learning and teachers' instruction more effective. Specifically, the first goal

of research studies is to help teachers make learning environments more **customized to contexts** (i.e., through instructional materials and technology, specifically pairing algorithms). I conclude that crowds can be a potential resource for helping time-strapped teachers revise and customize instructional materials and reveal the idiosyncratic preferences teachers may have for the materials. Additionally, the SimPairing study shows the practical consideration of customizing pairing policies and parameters in different classrooms. The second goal is to make learning environments more **differentiated** (i.e., by allowing the dynamic combination of different learning modes of individual and collaborative learning). The thesis studies demonstrate that teachers and students see the pedagogical value of such dynamic combinations, albeit I did not find a better learning outcome in the experiment compared to standard transitions. The third goal is to, through MMLA, make learning environments more **reflective** by aiding teachers to improve their teaching quality and teacher-student interaction.

While these characteristics are not the only ones conducive to creating effective learning environments, I believe these holds great potential for a more successful education ecology, which may allow more students to succeed at their own pace, and teachers to work more efficiently, facilitated by or collaborating with of AI-agents and systems, that are built to respect their preferences, boundaries and needs in teaching practices.

REFERENCE

- Alavi, H. S., & Dillenbourg, P. (2012). An Ambient Awareness Tool for Supporting Supervised Collaborative Problem Solving. *IEEE Transactions on Learning Technologies*, 5(3), 264–274.
- Aleahmad, T., Aleven, V., & Kraut, R. (2009). Creating a Corpus of Targeted Learning Resources with a Web-Based Open Authoring Tool. *IEEE Transactions on Learning Technologies*, 2(1), 3–9.
- Aleven, V., Borchers, C., Huang, Y., Nagashima, T., McLaren, B., Carvalho, P., & Koedinger, K. (2025). *An Integrated Platform for Studying Learning with Intelligent Tutoring Systems: CTAT+ TutorShop*. arXiv:2502.10395.
- Aleven, V., & Koedinger, K. R. (2013). Knowledge component (KC) approaches to learner modeling. *Design Recommendations for Intelligent Tutoring Systems*, 1, 165–182.
- Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, but only so Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, 26(1), 205–223.
- Alfonseca, E., Carro, R. M., Martín, E., Ortigosa, A., & Paredes, P. (2006). The impact of learning styles on student grouping for collaborative learning: a case study. *User Modeling and User-Adapted Interaction*, 16(3-4), 377–401.
- Alfredo, R., Mejia-Domenzain, P., Echeverria, V., Rahayu, D., Zhao, L., Alajlan, H., Swiecki, Z., Käser, T., Gašević, D., & Martinez-Maldonado, R. (2025). TeamTeachingViz: Benefits, challenges, and ethical considerations of using a multimodal analytics dashboard to support team teaching reflection. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 58–69.
- Almeda, M. V. Q. (2018). *When practice does not make perfect: Differentiating between productive and unproductive persistence*. Columbia University. <https://doi.org/10.7916/D83R291G>
- Alwahaby, H., Cukurova, M., Papamitsiou, Z., & Giannakos, M. (2022). The Evidence of Impact and Ethical Considerations of Multimodal Learning Analytics: A Systematic Literature Review. In M. Giannakos, D. Spikol, D. Di Mitri, K. Sharma, X. Ochoa, & R. Hammad (Eds.), *The Multimodal Learning Analytics Handbook* (pp. 289–325). Springer International Publishing.

- Amara, S., Macedo, J., Bendella, F., & Santos, A. (2016). Group Formation in Mobile Computer Supported Collaborative Learning Contexts: A Systematic Literature Review. *Journal of Educational Technology & Society*, 19(2), 258–273.
- Amershi, S., Weld, D., Vorvoreanu, M., & Fournery, A. (2019). Guidelines for human-AI interaction. *Proceedings of the*.
https://dl.acm.org/doi/abs/10.1145/3290605.3300233?casa_token=laDGIeyaLq0AAAAA:tFpJjLvoq9oSiujasZQSaRhtQAcZleG4NvZMBN6Q9DIZogY4VdXAh9DI9S66bapWFkIQoIqE79UL
- An, P., Bakker, S., Ordanovski, S., Taconis, R., & Eggen, B. (2018). ClassBeacons: Designing Distributed Visualization of Teachers' Physical Proximity in the Classroom. *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, 357–367.
- An, P., Bakker, S., Ordanovski, S., Taconis, R., Paffen, C. L. E., & Eggen, B. (2019). Unobtrusively Enhancing Reflection-in-Action of Teachers through Spatially Distributed Ambient Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. <https://doi.org/10.1145/3290605.3300321>
- An, P., Holstein, K., d'Anjou, B., Eggen, B., & Bakker, S. (2020). The TA Framework: Designing Real-time Teaching Augmentation for K-12 Classrooms. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–17.
- Aronson, E. (1978). *The Jigsaw Classroom*. 197. <https://psycnet.apa.org/fulltext/1980-51351-000.pdf>
- Bartel, Silla, & Vest. (n.d.). Reasoning about equations with tape diagrams: insights from math teachers and college students. *Proceedings of the Estonian Academy of Sciences. Biology, Ecology = Eesti Teaduste Akadeemia Toimetised. Bioloogia, Okoloogia*. <https://repository.isls.org/handle/1/7557>
- Bateman, B. (1992). Learning disabilities: the changing landscape. *Journal of Learning Disabilities*, 25(1), 29–36.
- Beardsley, M., Martínez Moreno, J., Vujovic, M., Santos, P., & Hernández-Leo, D. (2020). Enhancing consent forms to support participant decision making in multimodal learning data research. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, 51(5), 1631–1652.
- Beck, J. E., & Gong, Y. (2013). Wheel-Spinning: Students Who Fail to Master a Skill. *Artificial Intelligence in Education*, 431–440.

- Berg, K. F. (1994). Scripted Cooperation in High School Mathematics: Peer Interaction and Achievement. *Annual Meeting of the American Educational Research Association*. <https://eric.ed.gov/?id=ED374976>
- Bernacki, M. L., Greene, J. A., & Crompton, H. (2020). Mobile technology, learning, and achievement: Advances in understanding and measuring the role of mobile technology in education. *Contemporary Educational Psychology*, 60, 101827.
- Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2), 220–238.
- Blumenfeld, P., Fishman, B. J., Krajcik, J., Marx, R. W., & Soloway, E. (2000). Creating Usable Innovations in Systemic Reform: Scaling Up Technology-Embedded Project-Based Science in Urban Schools. *Educational Psychologist*, 35(3), 149–164.
- Borchers, C., Wang, Y., Karumbaiah, S., Ashiq, M., Shaffer, D. W., & Aleven, V. (2023). Revealing Networks: Understanding Effective Teacher Practices in AI-Supported Classrooms using Transmodal Ordered Network Analysis. In *arXiv [cs.CY]*. arXiv. <http://arxiv.org/abs/2312.10826>
- Boud, D., & Falchikov, N. (2007). *Rethinking Assessment in Higher Education: Learning for the Longer Term*. Routledge.
- Braun, V., & Clarke, V. (2016). (Mis)conceptualising themes, thematic analysis, and other problems with Fugard and Potts' (2015) sample-size tool for thematic analysis. *International Journal of Social Research Methodology*, 19(6), 739–743.
- Cao, J., Ganesh, A., Cai, J., Southwell, R., Perkoff, E. M., Regan, M., & D'Mello, S. (2023). A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 250–262.
- Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis – A general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems* (pp. 164–175). Springer Berlin Heidelberg.
- Chen, B., Hwang, G., & Lin, T. (2020). Impacts of a dynamic grouping strategy on students' learning effectiveness and experience value in an item bank-based collaborative practice system. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, 51(1), 36–52.

- Chen, Y., Mandel, T., Liu, Y.-E., & Popovic, Z. (2016). *Crowdsourcing Accurate and Creative Word Problems and Hints*. 12–21.
- Chen, Z., & Luo, B. (2014). Quasi-crowdsourcing testing for educational projects. *Companion Proceedings of the 36th International Conference on Software Engineering*. ICSE '14: 36th International Conference on Software Engineering, Hyderabad India. <https://doi.org/10.1145/2591062.2591153>
- Chuang, P.-J., Chiang, M.-C., Yang, C.-S., & Tsai, C.-W. (2012). Social networks-based adaptive pairing strategy for cooperative learning. *Journal of Educational Technology & Society*, 15(3), 226–239.
- Çimer, A., Sabiha, O. Ç., & Vekli, G. S. (2013). How does Reflection Help Teachers to Become Effective Teachers? *International Journal of Educational Research*, 1(4), 133–149.
- Conway, P. F. (2001). Anticipatory reflection while learning to teach: from a temporally truncated to a temporally distributed model of reflection in teacher education. *Teaching and Teacher Education*, 17(1), 89–106.
- Cornford, I. R. (2002). Reflective teaching: empirical research findings and some implications for teacher education. *Journal of Vocational Education & Training*, 54(2), 219–236.
- Corno, L. (2008). On Teaching Adaptively. *Educational Psychologist*, 43(3), 161–173.
- Cress, U., & Kimmerle, J. (2018). Collective knowledge construction. In *International handbook of the learning sciences* (pp. 137–146). Routledge.
- Cross, A., Bayyapunedi, M., Ravindran, D., Cutrell, E., & Thies, W. (2014, February 15). VidWiki. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW'14: Computer Supported Cooperative Work, Baltimore Maryland USA. <https://doi.org/10.1145/2531602.2531670>
- Cuendet, S., Bonnard, Q., Kaplan, F., & Dillenbourg, P. (2011). Paper interface design for classroom orchestration. *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 1993–1998.
- Cukurova, M., Giannakos, M., & Martinez-Maldonado, R. (2020). The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, 51(5), 1441–1449.
- Cukurova, M., Kent, C., & Luckin, R. (2019). Artificial intelligence and multimodal data in the service of

- human decision-making: A case study in debate tutoring. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, 50(6), 3032–3046.
- Davidoff, S., Lee, M. K., Dey, A. K., & Zimmerman, J. (2007). Rapidly Exploring Application Design Through Speed Dating. In *UbiComp 2007: Ubiquitous Computing* (pp. 429–446).
https://doi.org/10.1007/978-3-540-74853-3_25
- Day, E. A., Arthur, W., Bell, S. T., Edwards, B. D., Bennett, W., Mendoza, J. L., & Tubré, T. C. (2005). Ability-based pairing strategies in the team-based training of a complex skill: Does the intelligence of your training partner matter? *Intelligence*, 33(1), 39–65.
- Diana, N., Eagle, M., & Stamper, J. (2017). Automatic peer tutor matching: Data-driven methods to enable new opportunities for help. *Educational Data Mining*, 372–373.
- Diana, N., Eagle, M., Stamper, J., Grover, S., Bienkowski, M., & Basu, S. (2018). *Peer tutor matching for introductory programming: Data-driven methods to enable new opportunities for help*. 1377–1378.
- Dillenbourg, (2013). Design for classroom orchestration. *Computers & Education*.
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.837.6019&rep=rep1&type=pdf>
- Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. *Three Worlds of CSCL. Can We Support CSCL*, 61–91.
- Dillenbourg, P. A. (2002). *Over-scripting CSCL: The risks of blending collaborative learning with instructional design* (pp. 61–91).
- Dillenbourg, P., & Jermann, P. (2010). Technology for Classroom Orchestration. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education* (pp. 525–552). Springer New York.
- Di Mitri, D., Schneider, J., Trebing, K., Sopka, S., Specht, M., & Drachsler, H. (2020). Real-Time Multimodal Feedback with the CPR Tutor. *Artificial Intelligence in Education*, 141–152.
- Ding, N., Bosker, R. J., & Harskamp, E. G. (2011). Exploring gender and gender pairing in the knowledge elaboration processes of students using computer-supported collaborative learning. *Computers & Education*, 56(2), 325–336.
- Diziol, D., Rummel, N., Spada, H., & McLaren, B. M. (2007). *Promoting learning in mathematics: Script*

- support for collaborative problem solving with the cognitive tutor algebra*. Carnegie Mellon University.
<https://doi.org/10.1184/R1/6470348.V1>
- Dontcheva, M., Morris, R. R., Brandt, J. R., & Gerber, E. M. (2014). Combining crowdsourcing and learning to improve engagement and performance. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3379–3388.
- Dow, S., Gerber, E., & Wong, A. (2013). A pilot study of using crowds in the classroom. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 227–236.
- Drapeau, R., Chilton, L. B., Bragg, J., & Weld, D. S. (2016). MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 32–41.
- du Boulay, B. (2019). Escape from the Skinner Box: The case for contemporary intelligent learning environments. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, 50(6), 2902–2919.
- Echeverria, V., Holstein, K., Huang, J., Sewall, J., Rummel, N., & Aleven, V. (2020). Exploring Human–AI Control Over Dynamic Transitions Between Individual and Collaborative Learning. In *Addressing Global Challenges and Quality Education* (pp. 230–243). https://doi.org/10.1007/978-3-030-57717-9_17
- Echeverria, V., Yan, L., Zhao, L., Abel, S., Alfredo, R., Dix, S., Jaggard, H., Wotherspoon, R., Osborne, A., Shum, S. B., Gašević, D., & Martinez-Maldonado, R. (2024). TeamSlides: A multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space. *The 14th Learning Analytics and Knowledge Conference (LAK '24)*, 112–122.
- Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). *Education Technology: An Evidence-Based Review* (No. 23744). National Bureau of Economic Research. <https://doi.org/10.3386/w23744>
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Karns, K. (1998). High-Achieving Students' Interactions and Performance on Complex Mathematical Tasks as a Function of Homogeneous and Heterogeneous Pairings. In *American Educational Research Journal* (Vol. 35, Issue 2, pp. 227–267).
<https://doi.org/10.3102/00028312035002227>
- Giannakos, M., Spikol, D., Di Mitri, D., Sharma, K., Ochoa, X., & Hammad, R. (2022). *The Multimodal*

- Learning Analytics Handbook*. Springer Nature.
- Gibbons, S., & Farley, A. N. (2021). Learning to Think like a Teacher: Effects of Video Reflection on Preservice Teachers' Practice and Pedagogy. *Action in Teacher Education*, 43(3), 250–267.
- Glassman, E. L., Lin, A., Cai, C. J., & Miller, R. C. (2016). Learnersourcing personalized hints. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 27, 1626–1636.
- Grimes, K. J., & Stevens, D. D. (2009). Glass, Bug, Mud. *Phi Delta Kappan*, 90(9), 677–680.
- Håklev, S., Faucon, L., Hadzilacos, T., & Dillenbourg, P. (2017). Orchestration Graphs: Enabling Rich Social Pedagogical Scenarios in MOOCs. *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, 261–264.
- Hills, T. T. (2015). Crowdsourcing content creation in the classroom. *Journal of Computing in Higher Education*, 27(1), 47–67.
- Holstein, K., & Aleven, V. (2022). Designing for Human–AI Complementarity in K-12 Education. *AI Magazine*, 43(2), 239–248.
- Holstein, K., Aleven, V., & Rummel, N. (2020). A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. *Artificial Intelligence in Education*, 240–254.
- Holstein, K., Harpstead, E., Gulotta, R., & Forlizzi, J. (2020). Replay Enactments: Exploring Possible Futures through Historical Data. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (pp. 1607–1618). Association for Computing Machinery.
- Holstein, K., Hong, G., Tegene, M., McLaren, B. M., & Aleven, V. (2018). The classroom as a dashboard: co-designing wearable cognitive augmentation for K-12 teachers. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 79–88.
- Holstein, K., McLaren, B. M., & Aleven, V. (2017). Intelligent tutors as teachers' aides: exploring teacher needs for real-time analytics in blended classrooms. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 257–266.
- Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student Learning Benefits of a Mixed-Reality Teacher Awareness Tool in AI-Enhanced Classrooms. *Artificial Intelligence in Education*, 154–168.
- Holstein, K., McLaren, B. M., & Aleven, V. (2019a). Co-designing a real-time classroom orchestration tool to

- support teacher–ai complementarity. *Journal of Learning Analytics*, 6(2), 27–52.
- Holstein, K., McLaren, B. M., & Aleven, V. (2019b). Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. *Artificial Intelligence in Education*, 157–171.
- Holstein, K., & Olsen, J. K. (2021). Human–AI Co-orchestration: The Role of Artificial Intelligence in Orchestration. In *Handbook of Artificial Intelligence in Education*. Edward Elgar Publishing.
- Huang, Y.-M., Liao, Y.-W., Huang, S.-H., & Chen, H.-C. (2014). Jigsaw-based cooperative learning approach to improve learning outcomes for mobile situated learning. *Journal of Educational Technology & Society*, 17(1), 128–140.
- Huang, Y.-M., & Wu, T.-T. (2011). *A Systematic Approach for Learner Group Composition Utilizing U-Learning Portfolio*. *Educational Technology & Society*, Vol. 14.
- Hübscher, R. (2010). Assigning Students to Groups Using General and Context-Specific Criteria. *IEEE Transactions on Learning Technologies*, 3(3), 178–189.
- Javora, O., Děchtěrenko, F., Tetourová, T., Volná, K., & Brom, C. (2021). Customization in educational computer games and its effect on learning: Experimental study with primary school children. *Journal of Computer Assisted Learning*, 37(5), 1370–1382.
- Jessica Vitak, Yuting Liao, Anouk Mols, Daniel Trottier, Michael Zimmer, Priya C. Kumar, Jason Pridmore. (2023). When Do Data Collection and Use Become a Matter of Concern? A Cross-Cultural Comparison of U.S. and Dutch Privacy Attitudes. *International Journal of Communication Systems*, 17.
- Jiang, S., Yang, K., Suvarna, C., Casula, P., Zhang, M., & Rosé, C. (2019). Applying Rhetorical Structure Theory to Student Essays for Providing Automated Writing Feedback. *Proceedings of the NAACL Workshop on Discourse Relation Parsing and Treebanking*, 163–168.
- Jiang, Y., Schlagwein, D., & Benatallah, B. (2018). *A Review on Crowdsourcing for Education: State of the Art of Literature and Practice*. Conference: Pacific Asia Conference on Information Systems (PACIS). <https://aisel.aisnet.org/pacis2018/180>
- Johnson, D. W., & Johnson, R. T. (1987). *Learning together and alone: Cooperative, competitive, and individualistic learning*, 2nd ed. 2, 193.

- Kai, S., Almeda, M. V., Baker, R. S., Heffernan, C., & Heffernan, N. (2018). Decision Tree Modeling of Wheel-Spinning and Productive Persistence in Skill Builders. *Journal of Educational Data Mining*, 10(1), 36–71.
- Karumbaiah, S., Borchers, C., Falhs, A.-C., Holstein, K., Rummel, N., & Aleven, V. (2023, October 3). Teacher noticing and student learning in human-AI partnered classrooms: A multimodal analysis. *Proceedings of the 17th International Conference of the Learning Sciences - ICLS 2023*. 17th International Conference of the Learning Sciences (ICLS) 2023. <https://doi.org/10.22318/icls2023.151200>
- Karumbaiah, S., Borchers, C., Shou, T., Falhs, A. C., Liu, P., Nagashima, T., Rummel, N., & Aleven, V. (2023). A Spatiotemporal Analysis of Teacher Practices in Supporting Student Learning and Engagement in an AI-Enabled Classroom. *Artificial Intelligence in Education*, 450–462.
- Kerry, T., & Kerry, C. A. (1997). Differentiation: teachers' views of the usefulness of recommended strategies in helping the more able pupils in primary and secondary classrooms. In *Educational Studies* (Vol. 23, Issue 3, pp. 439–457). <https://doi.org/10.1080/0305569970230309>
- Kharrafa, A., Rix, S., Osadchiy, T., Preston, A., & Olivier, P. (2017). Group Spinner: Recognizing and Visualizing Learning in the Classroom for Reflection, Communication, and Planning. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5556–5567.
- Kim, D., & Lee, S. (2002). Designing Collaborative Reflection Supporting Tools in e-Project-Based Learning Environments. *J. Interact. Learn. Res.*, 13(4).
- Kirschner, P. A. (2015). Do we need teachers as designers of technology enhanced learning? *Instructional Science*, 43(2), 309–322.
- Kitto, K., Buckingham Shum, S., & Gibson, A. (2018). Embracing imperfection in learning analytics. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 451–460.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. In *Cognitive Science* (Vol. 36, Issue 5, pp. 757–798). <https://doi.org/10.1111/j.1551-6709.2012.01245.x>
- Korthagen, F. A. J. (2004). In search of the essence of a good teacher: towards a more holistic approach in teacher education. *Teaching and Teacher Education*, 20(1), 77–97.

- Kothiyal, A., Murthy, S., & Iyer, S. (2014). Think-pair-share in a large CS1 class: does learning really happen? *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education*, 51–56.
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, 86(1), 42–78.
- Kulkarni, A., Can, M., & Hartmann, B. (2012). Collaboratively crowdsourcing workflows with Turkomatic. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 1003–1012.
- Kulkarni, A., Narula, P., Rolnitzky, D., & Kontny, N. (2014). Wish: Amplifying Creative Ability with Expert Crowds. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2, 112–120.
- Kumar, L. (2015). Teaching and Learning. *Seminar Proceedings (Souvenir) of the NAAC Sponsored National Seminar on the Theme Teaching Strategies To Enhance Innovative Learning Process Organised by Loyola College of Education, Jamshedpur*, 27, 16–21.
- Kuo, Y.-C., Chu, H.-C., & Huang, C.-H. (2015). A learning style-based grouping collaborative learning approach to improve EFL students' performance in English courses. *Journal of Educational Technology & Society*, 18(2), 284–298.
- Lawrence, L., Guo, B., Yang, K., Echeverria, V., Kang, Z., Bathala, V., Li, C., Huang, W., Aleven, V., & Rummel, N. (2022). Co-designing AI-based orchestration tools to support dynamic transitions: Design narratives through conjecture mapping. *Computersupported Collaborative Learning*, 15.
<https://par.nsf.gov/biblio/10366101>
- Lin, C. Y., Huang, C. K., & Ko, C. J. (2020). The impact of perceived enjoyment on team effectiveness and individual learning in a blended learning business course: The mediating effect of knowledge sharing. *Australasian Journal of Educational*. <https://ajet.org.au/index.php/AJET/article/view/4446>
- Lin, Y.-T., Huang, Y.-M., & Cheng, S.-C. (2010). An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization. *Computers & Education*, 55(4), 1483–1493.
- Li, X., Yan, L., Zhao, L., Martinez-Maldonado, R., & Gasevic, D. (2023). CVPE: A Computer Vision Approach for Scalable and Privacy-Preserving Socio-spatial, Multimodal Learning Analytics. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 175–185.

- Lockheed, M. E., & Harris, A. M. (1984). Cross-Sex Collaborative Learning in Elementary Classrooms. In *American Educational Research Journal* (Vol. 21, Issue 2, pp. 275–294).
<https://doi.org/10.3102/00028312021002275>
- Long, Y., & Aleven, V. (2013). Supporting students' self-regulated learning with an open learner model in a linear equation tutor. *International Conference on Artificial Intelligence in Education*, 219–228.
- Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-Class Grouping: A Meta-Analysis. *Review of Educational Research*, 66(4), 423–458.
- Luther, K., Tolentino, J.-L., Wu, W., Pavel, A., Bailey, B. P., Agrawala, M., Hartmann, B., & Dow, S. P. (2015, February 28). Structuring, aggregating, and evaluating crowdsourced design critique. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15: Computer Supported Cooperative Work and Social Computing, Vancouver BC Canada.
<https://doi.org/10.1145/2675133.2675283>
- Manathunga, K., & Hernández-Leo, D. (2019). Flexible CSCL orchestration technology: Mechanisms for elasticity and dynamism in pyramid script flows. *International Society of the Learning Sciences (ISLS)*.
<https://repository.isls.org/handle/1/1574>
- Manathunga, K., Hernández-Leo, D., Caicedo, J., Ibarra, J. J., Martinez-Pabon, F., & Ramirez-Gonzalez, G. (2015). Collaborative Learning Orchestration Using Smart Displays and Personal Devices. *Design for Teaching and Learning in a Networked World*, 596–600.
- Maqtary, N., Mohsen, A., & Bechkoum, K. (2019). Group formation techniques in computer-supported collaborative learning: A systematic literature review. *Technology, Knowledge and Learning*, 24(2), 169–190.
- Marcos, J., Sanchez, E., & Tillema, H. (2011). Promoting teacher reflection: what is said to be done. *Journal of Education for Teaching*, 37(1), 21–36.
- Martinez-Maldonado, R., Clayphan, A., & Kay, J. (2015). Deploying and visualising teacher's scripts of small group activities in a multi-surface classroom ecology: A study in-the-wild. *Computer Supported Cooperative Work: CSCW: An International Journal*, 24(2-3), 177–221.
- Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2015). MTFeedback: Providing Notifications to

- Enhance Teacher Awareness of Small Group Work in the Classroom. *IEEE Transactions on Learning Technologies*, 8(2), 187–200.
- Martinez-Maldonado, R., Echeverria, V., Schulte, J., Shibani, A., Mangaroska, K., & Buckingham Shum, S. (2020). Moodoo: Indoor Positioning Analytics for Characterising Classroom Teaching. In *Lecture Notes in Computer Science* (pp. 360–373). Springer International Publishing.
- Martinez-Maldonado, R., Kay, J., Yacef, K., Edbauer, M.-T., & Dimitriadis, Y. (2013). *MTClassroom and MTDashboard: Supporting analysis of teacher attention in an orchestrated multi-tabletop classroom*. <https://doi.org/10.22318/cscl2013.1.320>
- Matsuda, N., Chandrasekaran, S., & Stamper, J. C. (2016). How quickly can wheel spinning be detected? *EDM*, 607–608.
- Matuk, C., Gerard, L., Lim-Breitbart, J., & Linn, M. (2016). Gathering Requirements for Teacher Tools: Strategies for Empowering Teachers Through Co-Design. *Journal of Science Teacher Education*, 27(1), 79–110.
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918.
- Mawasi, A., Ahmed, I., Walker, E., Wang, S., Marasli, Z., Whitehurst, A., & Wylie, R. (2020). *Using design-based research to improve peer help-giving in a middle school math classroom*. 1196.
- Means, B., Penuel, W. R., & Padilla, C. (2001). *The Connected School: Technology and Learning in High School*. Wiley.
- Mercier, E. (2016). Teacher orchestration and student learning during mathematics activities in a smart classroom. *International Journal of Smart Technology and Learning*, 1(1), 33–52.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2013). *Qualitative data analysis: A methods sourcebook*. Sage. *Thousand Oaks, CA. [Google Scholar]*.
- Mills, M., Monk, S., Keddle, A., Renshaw, P., Christie, P., Geelan, D., & Gowlett, C. (2014). Differentiated learning: from policy to classroom. *Oxford Review of Education*, 40(3), 331–348.
- Molenaar, I., Horvers, A., & Baker, R. S. (2019). Towards hybrid human-system regulation: Understanding children's SRL support needs in blended classrooms. *Proceedings of the 9th International Conference on*

- Learning Analytics & Knowledge*, 471–480.
- Molenaar, I., Horvers, A., Dijkstra, R., & Baker, R. S. (2020, March 23). Personalized visualizations to promote young learners' SRL. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. LAK '20: 10th International Conference on Learning Analytics and Knowledge, Frankfurt Germany. <https://doi.org/10.1145/3375462.3375465>
- Montrieux, H., Vanderlinde, R., Schellens, T., & De Marez, L. (2015). Teaching and Learning with Mobile Technology: A Qualitative Explorative Study about the Introduction of Tablet Devices in Secondary Education. *PloS One*, 10(12), e0144008.
- Morgan, H. (2014). Maximizing Student Success with Differentiated Learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 87(1), 34–38.
- Mujkanovic, A., Lowe, D., Willey, K., & Guetl, C. (2012). Unsupervised learning algorithm for adaptive group formation: Collaborative learning support in remotely accessible laboratories. *International Conference on Information Society (i-Society 2012)*, 50–57.
- Mullins, D., Rummel, N., & Spada, H. (2011). Are two heads always better than one? Differential effects of collaboration on students' computer-supported learning in mathematics. *International Journal of Computer-Supported Collaborative Learning*, 6(3), 421–443.
- Mu, T., Jetten, A., & Brunskill, E. (n.d.). Towards Suggesting Actionable Interventions for Wheel-Spinning Students. *The 13th International Conference on Educational Data Mining*, 183–193.
- Nagashima, T., Yang, K., Bartel, A., Silla, E., Vest, N., Alibali, M., & Alevan, V. (n.d.). Pedagogical Affordance Analysis: Leveraging teachers' pedagogical knowledge to elicit pedagogical affordances and constraints of instructional tools. *International Society of the Learning Sciences (2020)*.
- Niu, R., Jiang, L., & Deng, Y. (2018). Effect of Proficiency Pairing on L2 Learners' Language Learning and Scaffolding in Collaborative Writing. In *The Asia-Pacific Education Researcher* (Vol. 27, Issue 3, pp. 187–195). <https://doi.org/10.1007/s40299-018-0377-2>
- Nye, B. D. (2014). Barriers to ITS Adoption: A Systematic Mapping Study. *Intelligent Tutoring Systems*, 583–590.
- Olsen, J. (2017). *Orchestrating combined collaborative and individual learning in the classroom*.

<https://kilthub.cmu.edu/ndownloader/files/12255413>

- Olsen, J. K., Aleven, V., & Rummel, N. (2015). Toward Combining Individual and Collaborative Learning Within an Intelligent Tutoring System. *Artificial Intelligence in Education*, 848–851.
- Olsen, J. K., Rummel, N., & Aleven, V. (2019). It is not either or: An initial investigation into combining collaborative and individual learning using an ITS. *International Journal of Computer-Supported Collaborative Learning*, 14(3), 353–381.
- Olsen, J. K., Rummel, N., & Aleven, V. (2020). Designing for the co-Orchestration of Social Transitions between Individual, Small-Group and Whole-Class Learning in the Classroom. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-020-00228-w>
- O'Neill, S. C., & Stephenson, J. (2014). Evidence-based classroom and behaviour management content in Australian pre-service primary teachers' coursework: Wherefore art thou? *Australian Journal of Teacher Education*, 39(4). <https://doi.org/10.14221/ajte.2014v39n4.4>
- Ouhaichi, H., Spikol, D., & Vogel, B. (2023). Research trends in multimodal learning analytics: A systematic mapping study. *Computers and Education: Artificial Intelligence*, 4, 100136.
- Pang, Y., Xiao, F., Wang, H., & Xue, X. (2014). A Clustering-Based Grouping Model for Enhancing Collaborative Learning. *2014 13th International Conference on Machine Learning and Applications*, 562–567.
- Phillips, T. M., Saleh, A., & Ozogul, G. (2023). An AI toolkit to support teacher reflection. *International Journal of Artificial Intelligence in Education*, 33(3), 635–658.
- Pozdniakov, S., Martinez-Maldonado, R., Tsai, Y.-S., Cukurova, M., Bartindale, T., Chen, P., Marshall, H., Richardson, D., & Gasevic, D. (2022). The question-driven dashboard: How can we design analytics interfaces aligned to teachers' inquiry? *LAK22: 12th International Learning Analytics and Knowledge Conference*, 175–185.
- Prieto, L. P., Holenko Dlab, M., Gutiérrez, I., Abdulwahed, M., & Balid, W. (2011). Orchestrating technology enhanced learning: a literature review and a conceptual framework. *International Journal of Technology Enhanced Learning*, 3(6), 583–598.
- Prieto, L. P., Magnuson, P., Dillenbourg, P., & Saar, M. (2020). Reflection for action: designing tools to support

- teacher reflection on everyday evidence. *Technology, Pedagogy and Education*, 29(3), 279–295.
- Pugh, S. L., Rao, A., Stewart, A. E., & D'Mello, S. K. (2022). Do speech-based collaboration analytics generalize across task contexts? *International Learning Analytics and Knowledge Conference*, 208–218.
- Purcell, K., Heaps, A., Buchanan, J., & Friedrich, L. (2013). How teachers are using technology at home and in their classrooms. *Washington, DC: Pew Research Center's Internet & American Life Project*.
https://www.pewresearch.org/internet/wp-content/uploads/sites/9/media/Files/Reports/2013/PIP_TeachersandTechnologywithmethodology_PDF.pdf
- Quintana, R., Quintana, C., Madeira, C., & Slotta, J. D. (2016). Keeping Watch: Exploring Wearable Technology Designs for K-12 Teachers. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2272–2278.
- Raca, M., & Dillenbourg, P. (2013). System for assessing classroom attention. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 265–269.
- Retnowati, E., Ayres, P., & Sweller, J. (2017). Can collaborative learning improve the effectiveness of worked examples in learning mathematics? *Journal of Educational Psychology*, 109(5), 666–679.
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249–255.
- Rodríguez-Triana, M. J., Prieto, L. P., Martínez-Monés, A., Asensio-Pérez, J. I., & Dimitriadis, Y. (2018). *The Teacher in the Loop: Customizing Multimodal Learning Analytics for Blended Learning*. 417–426.
- Romano, M., & Schwartz, J. (2005). Exploring technology as a tool for eliciting and encouraging beginning teacher reflection. *Contemporary Issues in Technology and Teacher Education*, 5(2), 149–168.
- Saar, M., Prieto, L. P., Rodríguez-Triana, M. J., & Kusmin, M. (2018, July). Personalized, teacher-driven in-action data collection: Technology design principles. *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), Mumbai. <https://doi.org/10.1109/icalt.2018.00020>
- Sandmire, D. A., & Boyce, P. F. (2004). Pairing of opposite learning styles among allied health students: effects on collaborative performance. *Journal of Allied Health*, 33(2), 156–163.
- Sankaranarayanan, S., Reddy, K. S., Cao, M., Maronna, I., An, H., Bogart, C., Murray, R. C., Hilton, M., Sakr,

- M., & Carolyn, P. R. (2020). Designing for learning during collaborative projects online: tools and takeaways. *Information and Learning Sciences*, 121(7/8), 569–577.
- Santos, L. P. P. (2012). *Supporting orchestration of blended CSCL scenarios in Distributed Learning Environments*. <https://pdfs.semanticscholar.org/c135/fbc5605a26505d0f0b69278e1bf46886d242.pdf>
- Schuwert, R., & Kusters, R. (2014). Mass customization of education by an institution of HE: What can we learn from industry? *The International Review of Research in Open and Distributed Learning*, 15(2). <https://doi.org/10.19173/irrodl.v15i2.1704>
- Schwartz, H. S., & Schon, D. A. (1987). The reflective practitioner: How professionals think in action. *Administrative Science Quarterly*, 32(4), 614.
- Sergis, S., & Sampson, D. G. (2017). Teaching and Learning Analytics to Support Teacher Inquiry: A Systematic Literature Review. In A. Peña-Ayala (Ed.), *Learning Analytics: Fundamentals, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning* (pp. 25–63). Springer International Publishing.
- Sharma, S. K., Palvia, S. C. J., & Kumar, K. (2017). Changing the landscape of higher education: From standardized learning to customized learning. *Journal of Information Technology Case and Application Research*, 19(2), 75–80.
- Sharples, M. (2013). Shared orchestration within and beyond the classroom. *Computers & Education*, 69, 504–506.
- Shou, T., Borchers, C., Karumbaiah, S., & Aleven, V. (2023). Optimizing Parameters for Accurate Position Data Mining in Diverse Classrooms Layouts. *International Educational Data Mining Society*. <https://eric.ed.gov/?id=ED630870>
- Singh, P. (2008). Oral and written reflection in practice management: An action research approach. *Systemic Practice and Action Research*, 21(2008), 171–185.
- Smirani, L. K., Yamani, H. A., Menzli, L. J., & Boulahia, J. A. (2022). Using ensemble learning algorithms to predict student failure and enabling customized educational paths. *Scientific Programming*, 2022, 1–15.
- Sokolov, M. (2001). Technology's impact on society: The issue of mass-customized education. *Technological Forecasting and Social Change*, 68(2), 195–206.

- Solemon, B., Ariffin, I., Din, M., & Anwar, R. (2013). A Review of the uses of Crowdsourcing in Higher Education. *International Journal of Asian Social Science*, 3(9), 2066–2073.
- Souvignier, E., & Kronenberger, J. (2007). Cooperative learning in third graders' jigsaw groups for mathematics and science with and without questioning training. *The British Journal of Educational Psychology*, 77(Pt 4), 755–771.
- Srba, I., & Bielikova, M. (2015). Dynamic Group Formation as an Approach to Collaborative Learning Support. *IEEE Transactions on Learning Technologies*, 8(2), 173–186.
- Storch, N., & Aldosari, A. (2013). Pairing learners in pair work activity. *Language Teaching Research*, 17(1), 31–48.
- Sutherland, J. A., & Topping, K. J. (1999). Collaborative creative writing in eight-year-Olds: Comparing cross-ability fixed role and same-ability reciprocal role pairing. *Journal of Research in Reading*, 22(2), 154–179.
- Tamim, H. M., Sultana, F., Tasneem, N., Marzan, Y., & Khan, M. M. (2021). Class Insight: A Student Monitoring System with Real-time Updates using Face Detection and Eye Tracking. *2021 IEEE World AI IoT Congress (AIIoT)*, 0213–0220.
- Taniguchi, Y., Gao, Y., Kojima, K., & Konomi, S. (2018). Evaluating Learning Style-Based Grouping Strategies in Real-World Collaborative Learning Environment. In *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts* (pp. 227–239). https://doi.org/10.1007/978-3-319-91131-1_18
- Tien, H.-W., Lin, Y.-S., Chang, Y.-C., & Chu, C.-P. (2013). A genetic algorithm-based multiple characteristics grouping strategy for collaborative learning. *International Conference on Web-Based Learning*, 11–22.
- Tomlinson, C. A. (2014). *The Differentiated Classroom: Responding to the Needs of All Learners*. ASCD.
- van Leeuwen, A., Rummel, N., Holstein, K., McLaren, B. M., Aleven, V., Molenaar, I., Campen, C. K., Schwarz, B., Prusak, N., Swidan, O., & Others. (2018). *Orchestration tools for teachers in the context of individual and collaborative learning: what information do teachers need and what do they do with it?* International Society of the Learning Sciences, Inc.[ISLS].
- VanLehn, K., Burkhardt, H., Cheema, S., Kang, S., Pead, D., Schoenfeld, A., & Wetzel, J. (2018). The Effect of Digital Versus Traditional Orchestration on Collaboration in Small Groups. *Artificial Intelligence in*

- Education*, 369–373.
- Vaughn, S., & Linan-Thompson, S. (2003). What Is Special About Special Education for Students with Learning Disabilities? *The Journal of Special Education*, 37(3), 140–147.
- Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *The American Behavioral Scientist*, 57(10), 1500–1509.
- Waalkens, M., Aleven, V., & Taatgen, N. (2013). Does supporting multiple student strategies lead to greater learning and motivation? Investigating a source of complexity in the architecture of intelligent tutoring systems. In *Computers & Education* (Vol. 60, Issue 1, pp. 159–171).
<https://doi.org/10.1016/j.compedu.2012.07.016>
- Walker, E., Rummel, N., & Koedinger, K. R. (2014). Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. *International Journal of Artificial Intelligence in Education*, 24(1), 33–61.
- Wang, P., Tchounikine, P., & Quignard, M. (2018). Chao: a framework for the development of orchestration technologies for technology-enhanced learning activities using tablets in classrooms. *International Journal of Technology Enhanced Learning*, 10(1-2), 1–21.
- Wang, X., Thompson, M., Yang, K., Roy, D., Koedinger, K. R., Rose, C. P., & Reich, J. (2021). Practice-Based Teacher Questioning Strategy Training with ELK: A Role-Playing Simulation for Eliciting Learner Knowledge. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), 1–27.
- Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research*, 13(1), 21–39.
- Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity Issues in Collaborative Group Assessment: Group Composition and Performance. *American Educational Research Journal*, 35(4), 607–651.
- Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. C. Berliner (Ed.), *Handbook of educational psychology*, (pp (Vol. 1071, pp. 841–873). Macmillan Library Reference Usa; London, England.
- Whitehill, J., & Seltzer, M. (2017). *A Crowdsourcing Approach to Collecting Tutorial Videos-Toward Personalized Learning-at-Scale*. 157–160.

- Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., & Heffernan, N. (2016).
 AXIS: Generating explanations at scale with learnersourcing and machine learning. *L@S 2016 - Proceedings of the 3rd 2016 ACM Conference on Learning at Scale*, 379–388.
- Wise, A. F., & Jung, Y. (2019). Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, 6(2), 53–69.
- Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, 140, 103599.
- Yang, K. B., Borchers, C., Falhs, A.-C., Echeverria, V., Karumbaiah, S., Rummel, N., & Aleven, V. (2024). Leveraging multimodal classroom data for teacher reflection: Teachers’ preferences, practices, and privacy considerations. *European Conference on Technology Enhanced Learning (EC-TEL)*, 498–511.
- Yang, K. B., Echeverria, V., Wang, X., Lawrence, L., Holstein, K., Rummel, N., & Aleven, V. (2021a). Exploring Policies for Dynamically Teaming up Students through Log Data Simulation. *International Educational Data Mining Society*. <https://eric.ed.gov/?id=ED615570>
- Yang, K. B., Echeverria, V., Wang, X., Lawrence, L., Holstein, K., Rummel, N., & Aleven, V. (2021b). Exploring Policies for Dynamically Teaming Up Students through Log Data Simulation. *Educational Data Mining*.
- Yang, K. B., Lawrence, L., Echeverria, V., Guo, B., Holstein, K., Rummel, N., & Aleven, V. (n.d.). “I like student choice, program insights, but final say from the teacher”: Teachers’ Preferences regarding Human-AI Control in Dynamic Student Pairing. *European Conference of Technology Enhanced Learning*.
- Yang, K. B., Lawrence, L., Echeverria, V., Guo, B., Rummel, N., & Aleven, V. (2021). Surveying teachers’ preferences and boundaries regarding human-AI control in dynamic pairing of students for collaborative learning. In *Technology-Enhanced Learning for a Free, Safe, and Sustainable World* (pp. 260–274). Springer International Publishing.
- Yang, K. B., Nagashima, T., Yao, J., Williams, J. J., Holstein, K., & Aleven, V. (2021). Can Crowds Customize Instructional Materials with Minimal Expert Guidance? Exploring Teacher-guided Crowdsourcing for Improving Hints in an AI-based Tutor. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), 1–24.

- Yan, L., Zhao, L., Gasevic, D., & Martinez-Maldonado, R. (2022). Scalability, Sustainability, and Ethicality of Multimodal Learning Analytics. *LAK22: 12th International Learning Analytics and Knowledge Conference*, 13–23.
- Zhang, C., Huang, Y., Wang, J., Lu, D., Fang, W., Stamper, J., Fancsali, S., Holstein, K., & Aleven, V. (2019). Early Detection of Wheel Spinning: Comparison across Tutors, Models, Features, and Operationalizations. *International Educational Data Mining Society*. <https://eric.ed.gov/?id=ED599222>